Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

The Interplay of Achievement and Achievement Motivation: Gender Differences in Math Top-Performers and Functional Relations

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

Vorgelegt von

Lena Kristina Keller

Master of Science

Berlin, 2020

Erstgutachter:

Prof. Dr. Martin Brunner, Universität Potsdam/Freie Universität Berlin

Zweitgutachterin:

Prof. Dr. Franzis Preckel, Universität Trier

Tag der Disputation: 31.07.2020

Acknowledgement

First and foremost, I would like to express my deepest gratitude to Martin Brunner for his guidance and support throughout this dissertation project, for triggering my passion for graphics, and the trust, patience, and freedom he gave me. I am also deeply grateful to Franzis Preckel for her continuous guidance and supervision, which already started at my time as a student research assistant at the University of Trier. With you, I had the best supervisor team – thank you both for sharing your immense knowledge and inspiring research enthusiasm with me!

Furthermore, I wish to thank Prof. Dr. Ursula Kessels, Prof. Dr. Steffi Pohl, and Dr. Georg Hosoya for their interest, time, and effort put into reading my doctoral thesis and their willingness to be part of my dissertation defense committee.

My thanks also go to the International Max Planck Research School on the Life Course (LIFE) that made this endeavor possible by supporting my dissertation project financially with a stipend and scientifically with the opportunity to discuss and exchange ideas in interdisciplinary and international environment in seminars, academies, and workshops. I enjoyed this LIFE a lot!

Special thanks go to Prof. Dr. Kai S. Cortina, Prof. Dr. Jacque S. Eccles, and Dr. Leonie Kronborg for giving me the opportunity to visit their labs at the University of Michigan, UC Irvine, and Monash University. Thank you for all the things I learned from you. You broadened my horizon.

In addition, I would like to thank my current and former colleagues at the University of Potsdam and the Freie Universität Berlin who made this journey so much more pleasant: Anna, Andrea, Anke, Anta, Gesine, Julia, Kathi, Marina, Kasia, Katharina,

Sophie, and our student research assistants. Many thanks also to all members of the ISQ who welcomed me so warmly when I started my PhD in 2015 at the FU.

Furthermore, I am utterly grateful to Veit Kubik for his critical, thorough, and constructive feedback on this thesis, and Andrea Hasl for her attentive and helpful comments on an earlier draft.

Moreover, I would like to thank my family and friends for all their unconditional support and encouragement. Thank you for believing in me, for supporting me, and for distracting me when I needed it.

Most importantly, I wish to thank Veit for being the most loving, supportive, and inspiring partner I could ever wish for. Thank you for being such a “smart cookie” and my only possible soulmate.

Summary

Achievement and achievement motivation are two central constructs in educational psychology. The interplay between these constructs is a major element in prominent theoretical frameworks such as the Situated Expectancy Value Model (SEVT; e.g., Eccles & Wigfield, 2020). Given the topic's relevance for individuals' further educational trajectories and life courses, it is important to obtain particularly reliable and robust results. To achieve this aim in the present doctoral thesis, I applied innovative multilevel meta-analytical approaches, using data from international large-scale assessments to investigate the interplay between achievement and achievement motivation.

Using a *Multilevel Individual Participant Data (IPD) Meta-Analysis*, Study I examined gender differences in achievement, achievement profiles, and achievement motivation in mathematics, reading, and science in the group of top-performing math students (top 5%) across 82 countries. In addition, it was investigated to what extent gender differences in the top 5% in mathematics were moderated by cross-national variations in sociocultural factors (i.e., in specific gender equality indicators). To this end, I used data from 15-year-old students who participated in six PISA cycles. The results showed that there were on average more male than female students (40%) that scored in the top 5% in mathematics. In addition, mathematically top-performing female students' achievement profiles were more balanced across domains, whereas mathematically top-performing male students' achievement profiles were more mathematics-oriented. Moreover, mathematically top-performing female students reported a higher interest in the verbal domain and in human biology than male students. On the contrary, mathematically top-performing male students reported a higher interest in physics-related topics than female students (i.e., physics, motion of forces, energy transformation). The results also

showed that specific gender equality indicators moderated the share of female students in the top 5% in mathematics and explained variability in achievement profiles.

In Study II of this doctoral thesis, the functional relations between achievement and self-concept were systematically investigated using a *Multilevel Integrative Data Analysis*. The guiding research question was to examine the extent to which a nonlinear relation between achievement and self-concept can be generalized across domains, age groups, analytical approaches, and 13 countries. The analyses were based on eight cycles of PISA, TIMSS, and PIRLS. Quadratic and interrupted regression analyses showed nonlinear relations in secondary school students, demonstrating that the relations between achievement and corresponding self-concepts were weaker for lower achieving students than for higher achieving students. This suggests that lower achieving students might apply self-protective strategies to prevent negative self-evaluation. Nonlinear effects were also present in younger students, but the pattern of results was rather heterogeneous.

The present doctoral thesis contributed to uncover the interplay between achievement and achievement motivation by using advanced multilevel meta-analytical approaches. Based on this work, future research is encouraged to apply such statistical tools to meta-analyze variance in individual participant data to enhance the reliability and robustness of the obtained empirical evidence on the interplay between achievement and achievement motivation.

Zusammenfassung

Leistung und Leistungsmotivation sind zentrale Konstrukte in der pädagogisch-psychologischen Forschung. Das Zusammenspiel von Leistung und Leistungsmotivation wird in prominenten theoretischen Rahmenmodellen wie der Erwartungs-Wert-Theorie (z.B. Eccles & Wigfield, 2020) untersucht. Aufgrund der zentralen Bedeutung dieses Themas für individuelle Bildungs- und Lebensverläufe, ist es wichtig reliable und robuste Ergebnisse zu gewinnen. Um dieses Ziel zu erreichen, wurden in der vorliegenden Dissertation innovative meta-analytische Ansätze und Daten internationaler Schulleistungsstudien verwendet, um das Zusammenspiel von Leistung und Leistungsmotivation zu untersuchen.

Im Rahmen einer *Multilevel Individual Participant Data (IPD) Meta-Analyse* wurden in Teilstudie I Geschlechtsunterschiede in der Leistung, in Leistungsprofilen und in der Leistungsmotivation von mathematisch talentierten Schülerinnen und Schüler (Top 5%) in Mathematik, im Lesen und in Naturwissenschaften in 82 Ländern analysiert. Zudem wurde untersucht, inwiefern die Variation in den Geschlechtsunterschieden zwischen Ländern auf Unterschiede in soziokulturellen Faktoren (d.h., in spezifischen Indikatoren der Geschlechtergleichstellung) zurückzuführen ist. Hierfür wurden die Daten von 15-jährigen Mädchen und Jungen aus sechs PISA-Zyklen verwendet. Es konnte gezeigt werden, dass insgesamt weniger Mädchen in der Gruppe der Spitzenleistenden in Mathematik (Top 5%) vertreten waren (Mädchenanteil 40%) sowie Mädchen in dieser Gruppe balanciertere Leistungsprofile aufwiesen, während Jungen eher zu mathematik-orientierten Leistungsprofilen neigten. Weiterhin zeigte sich, dass mathematisch talentierte Mädchen eine höhere Motivation im verbalen Bereich sowie ein stärkeres Interesse an Humanbiologie als Jungen berichteten. Mathematisch talentierte Jungen berichteten

hingegen ein größeres Interesse an den Themenbereichen Physik, Bewegung und Kräfte und Energieumwandlung als Mädchen. Zudem konnte in Teilstudie I gezeigt werden, dass spezifische Gleichstellungsindikatoren den Anteil der Schülerinnen in den Top 5% in Mathematik moderierten und Variabilität in den Leistungsprofilen erklärten.

In Teilstudie II wurde der funktionale Zusammenhang zwischen Leistung und Selbstkonzept systematisch im Rahmen einer *Multilevel Integrative Data Analysis* untersucht. Die zentrale Forschungsfrage war, ob ein nicht-linearer Zusammenhang zwischen Leistung und Selbstkonzept über Inhaltsdomänen, Altersgruppen, Analysemethoden und 13 Länder hinweg generalisiert vorliegt. Die Analysen basierten auf acht Zyklen der PISA-, TIMSS- und PIRLS-Studien. Die Ergebnisse zeigten, dass nicht-lineare Zusammenhänge zwischen Leistung und korrespondierenden Selbstkonzepten in Mathematik und im verbalen Bereich bei Schülerinnen und Schülern der Sekundarstufe vorlagen. Dabei deuten die Ergebnisse der quadratischen Regressionen und der *Interrupted Regressions* darauf hin, dass der Zusammenhang für leistungsschwächere Schülerinnen und Schüler schwächer war als für leistungsstärkere Schülerinnen und Schüler. Dies könnte in der Anwendung selbstwertdienlicher Strategien begründet sein. Nicht-lineare Zusammenhänge zeigten sich auch für jüngere Schülerinnen und Schüler, jedoch war die Befundlage für diese Altersgruppe über Länder und Analysemethoden hinweg heterogener.

Die vorliegende Doktorarbeit trägt mit diesen Erkenntnissen dazu bei, das Zusammenspiel von Leistung und Leistungsmotivation unter Anwendung von *IPD-Meta-Analysen* bzw. *Integrativen Datenanalysen* aufzuklären. Basierend auf dieser Arbeit wird die Bedeutung hervorgehoben, Daten auf der individuellen Personenebene zu meta-analysieren, um die Reliabilität und Robustheit von Befunden zum Zusammenspiel von Leistung und Leistungsmotivation zu erhöhen.

Table of Contents

1

Introduction and Theoretical Background

1 Introduction and Theoretical Background

There are numerous factors that influence individuals' educational and occupational choices. Among them, students' achievement and achievement motivation are of special importance (e.g., Eccles & Wigfield, 2020). International large-scale assessments, such as the Programme for International Student Assessment (PISA), have started to shift their attention from an exclusive focus on students' achievement to an integrative view of students' achievement and achievement motivation to be important for individuals' successful participation in society (OECD, 2015).

For example, two high school students, Toby and Tina, will very likely choose different college majors. Tina performs very well in all subjects, but is especially interested in biology and dreams of becoming a physician to be able to help people. On the other hand, Toby excels in mathematics and physics. Although he could do better in the other subjects at school, he is not much interested in them. He thinks of himself as a math person and wants to become an engineer and develop new technologies for energy transformations later in life. This example illustrates that there are different facets of achievement motivation (e.g., academic self-concepts and interest), that achievement and beliefs about one's own academic abilities (i.e., academic self-concepts) are closely related, and that achievement and achievement motivation can differ depending on the content domain. Furthermore, the case of Toby and Tina indicates that achievement and achievement motivation might differ for female and male students. However, students do not develop in a vacuum, but in environments that provide them with opportunities and also set expectations. These expectations and opportunities vary for female and male students and likely influence their achievement and achievement motivation (e.g., Baker & Jones, 1993; Eccles, 1994; Else-Quest et al., 2010).

Importantly, scientific research is currently facing concerns about the replicability of results in many areas of psychology, education, and other fields (e.g., Ioannidis, 2005; Open Science Collaboration, 2015). One way to advance psychology as a field might be to apply research synthesis methods (Roisman & van IJzendoorn, 2018). Research syntheses summarize the results of single studies by using meta-analytical techniques. With these techniques, weighted average effect sizes, variations in effect sizes between studies, and factors that moderate the size of the effects can be estimated and investigated (Shadish et al., 2002). As a result, research synthesis fosters reproducible, rigorous, and transparent research (McNutt, 2014). Two relatively new forms of meta-analysis are individual participant data (IPD) meta-analyses and integrative data analyses that add the level of the participants to the analyses (e.g., Cooper & Patall, 2009; Curran & Hussong, 2009). These new forms have several advantages compared with traditional meta-analyses (e.g., Reily et al., 2010).

In this doctoral thesis, I aim to investigate the interplay between achievement and achievement motivation in two studies by following two research strands that provide different angles on the interplay between achievement and achievement motivation. These research strands will be discussed within the framework of the Situated Expectancy–Value Theory (SEVT) of achievement performance and choice (e.g., Eccles et al., 1983; Eccles & Wigfield, 2020). In the *first research strand,* which I cover in Study I, I investigated the extent to which gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in mathematics, reading, and science across countries exist (Research Question 1). Furthermore, I examined to what extent cross-national gender differences in the group of top-performing math students were related to sociocultural factors, or more specifically, to the level of gender equality in a country (Research Question 2). To do so, I performed a multilevel IPD meta-analysis that

synthesized gender differences in the top 5% in mathematics by using data from six cycles from the Programme for International Students Assessment (PISA 2000–2015, 15-year-olds, 82 countries). Furthermore, I conducted multivariate moderator analyses using specific gender equality indicators by the United Nations (UN), the Organisation for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organisation (UNESCO), and the International Labour Organization (ILO).

The *second research strand*, which I address in Study II, refers to the question of how achievement and academic self-concept—a central motivational construct in educational psychology—are functionally related (Research Question 3). To tackle this research question, I investigated whether nonlinear relations between these constructs can be generalized across domains (mathematics and the verbal domain), age groups (elementary and secondary school students), analytical methods (polynomial and interrupted regression), and 13 countries. To do so, I performed an integrative data analysis and synthesized data from eight cycles from the Trends in Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and PISA.

In the following chapters, I first present the theoretical background (Chapter 1) in which the definitions, measurement, and relevance of achievement and achievement motivation will be presented (Section 1.1). Then, I discuss how selected theories address the interplay between achievement and achievement motivation with an emphasis on the SEVT as the guiding theoretical framework of this thesis (Section 1.2.2). Based on these prior sections, I introduce the topic of the first research strand of the present thesis, which refers to gender differences in education and educational trajectories in top-performing math students; the terms *gender* and *sex* are terminologically defined in this dissertation, an overview is provided of research on gender differences in the general population and in top-performing

math students, and gender differences are explained in terms of the SEVT (Section 1.3.4). In the next section, I introduce the topic of the second research strand of this doctoral thesis, which refers to the relation of achievement and academic self-concept, and how this relation is stated within the SEVT (Section 1.4). Before describing the objectives of the present doctoral thesis (Section 1.6), the two methodological approaches are presented for studying the interplay between achievement and achievement motivation; they refer to approaches modelling intraindividual hierarchies of achievement and achievement motivation as well as research synthesis methods (Section 1.5).

The theoretical background is followed by Study I (Chapter 2) and Study II (Chapter 3) and the General Discussion (Chapter 4). In the General Discussion, the main results related to the research questions are briefly summarized and discussed in relation to the topics that were introduced in the theoretical background and beyond (Sections 4.1–4.3). Finally, strengths, limitations, and directions for further research (Section 4.4) as well as implications for (educational) policy and practice are discussed (Section 4.5).

1.1 Achievement and Achievement Motivation: Definitions, Measurement, and Relevance

Fostering students' achievement and achievement motivation is a central educational goal in school curricula worldwide (OECD, 2015; World Economic Forum, 2015). There is a global consensus that a focus on both the skills and knowledge of students as well as their motivational tendencies is necessary to enable them to live fulfilled lives, meet challenges, and make the most of the opportunities available to them (Bertling et al., 2016; Schunk & Mullen, 2013).

Student *achievement* is assumed to be the result of a long-term, cumulative, and domain-specific process of knowledge acquisition (Baumert et al., 2009). Achievement is

not only an outcome of learning, but also the result of prior achievement, which pays off in the speed, ease, and quality of continued learning (Baumert et al., 2009). Student achievement can be measured by grades provided by teachers or standardized tests. Compared to grades, standardized tests have the advantage to be an objective measure of student achievement. This renders student achievement to be directly comparable across classes, schools, and even *across* countries (Brookhart, 2015).

Motivation is assumed to energize and direct individuals' actions (Pintrich, 2003). In achievement-related contexts, motivation is related to motives that let students act in different situations, to beliefs why and how they are doing what they are doing, as well as to students' decisions about how to direct their behaviors (Pintrich, 2003; Weiner, 1992; Wigfield et al., 2006, 2020). To this end, *achievement motivation* comprises different types and qualities of motivation, including needs, drives, goals, aspirations, interests, and affects (Lazowski & Hulleman, 2016). Similar to achievement, achievement motivation can be considered as an outcome of learning (e.g., students' motivation increases or decreases as a consequence of learning successes or failures), but also as a prerequisite for future learning because motivated students are more persistent, learn more, and show a higher level of elaboration (Lazowski & Hulleman, 2016; Pintrich, 2003; Wigfield et al., 2006). Student achievement motivation can be measured in several ways. Most frequently, they are assessed through self-reports, but also alternative approaches to the measurement of motivation exist, including phenomenological, neuropsychological, and behavioral approaches (Duckworth & Yeager, 2015; Fulmer & Frijters, 2009).

Assessing both constructs in standardized ways across nations and age-groups is critical. International large-scale assessments in education provide representative, high-quality data on students' achievement and achievement motivation. The largest and most prominent international large-scale assessments are the PISA (conducted since 2000 by the

OECD), TIMSS, and PIRLS (conducted since 1995 and 2001 by the International Association for the Evaluation of Educational Achievement [IEA]; Kirsch et al., 2012). In the present doctoral thesis, data from PISA, TIMSS, and PIRLS are used to investigate the interplay between achievement (as assessed by standardized tests) and achievement motivation (as measured by self-reports).

1.2 Interplay Between Achievement and Achievement Motivation

One of the major goals in educational psychology is to explain achievement-related performance and choices, specifically in educational and occupational contexts. To achieve this, prior work has largely studied the interplay of multiple achievement- and motivation-related factors, such as basic cognitive abilities, domain-specific knowledge, interests, preferences, or personality traits. It is assumed that the interplay of achievement and achievement motivation plays a critical role in students' learning (Bast & Reitsma, 1998), for example, in terms of the amount of knowledge or expertise accumulated in specific areas.

1.2.1 General Theories

Given the prominent interest in this research field, various theories have been developed that address the interplay between achievement and achievement motivation. Among many others, these include the (extended) Theory of Work Adjustment (Dawis & Lofquist, 1984; Lubinski & Benbow, 2006) and the PPIK Theory (intelligence-as-Process, Personality, Interests, and intelligence-as-Knowledge; Ackerman, 1996). For example, the (*extended) Theory of Work Adjustment* takes into account the interplay among individuals' abilities, preferences, and interests as well as the ability requirements and rewards from the academic or occupational environments to explain educational commitment and

occupational tenure (Dawis & Lofquist, 1984; Lubinski & Benbow, 2006). Alternatively, the *PPIK Theory* concentrates on the interplay among basic cognitive skills, personality traits, interests, and domain-specific knowledge to explain individuals' successful accumulation of knowledge or expertise in specific areas (Ackerman, 1996). Although many of these theories can be applied to explain achievement-related performance and choices in general, they are not specifically tailored to educational contexts and thus lack specificity to guide further theory development and empirical investigation in this specific field of research.

Of more specific relevance is the *Situated Expectancy–Value Theory of achievement performance and choice* (SEVT; formerly expectancy–value theory of achievement-related choices; e.g., Eccles et al., 1983; Eccles & Wigfield, 2020). The SEVT specifically focusses on the interplay between educationally relevant aspects of achievement and achievement motivation (Eccles et al., 1983). To this end, I will present the SEVT in greater detail in the following section.

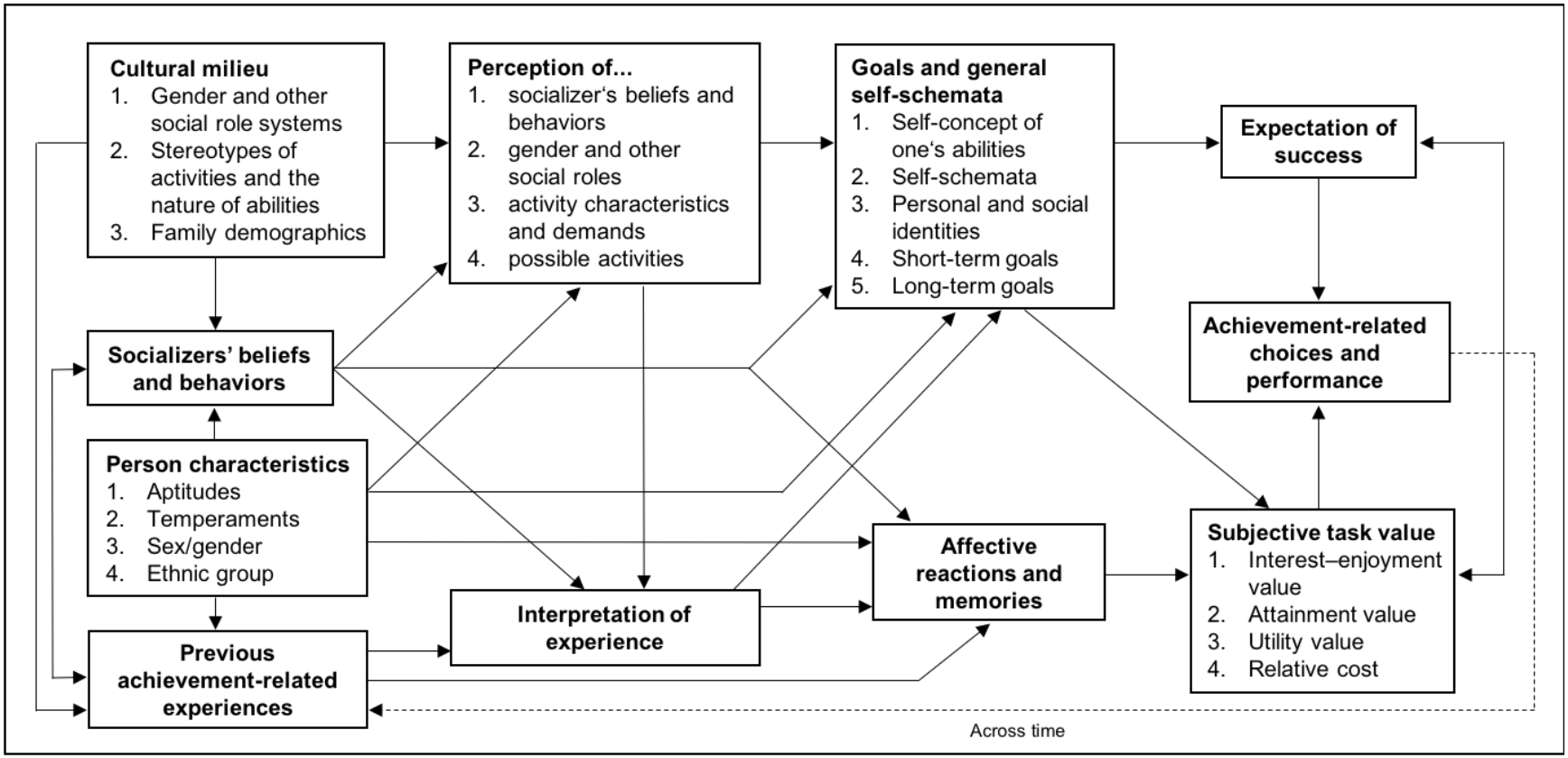
1.2.2 Situated Expectancy–Value Theory (SEVT)

Based on the SEVT, individuals' achievement and achievement-choices are most closely related to individuals' expectancies of success (i.e., confidence to succeed in a task) and individuals' subjective task values (i.e., the perceived interest, usefulness, importance, and cost of a task; e.g., Eccles, 1983; Eccles & Wigfield, 2020; Wigfield & Eccles, 2020). For example, considering expectancies of success, a student with a low self-concept in mathematics likely believes not to be good in math-related tasks. This lets him or her expect to perform poorly in a specific math exam, which in turn negatively affects the student's performance. Similarly, but related to individuals' task values, there is a student with the intention to become a natural scientist; he or she ascribes mathematics a high

utility value and even a high interest–enjoyment value and therefore is strongly motivated to prepare for the math exam, which in turn positively affects the student's performance. Together, individuals' expectancies of success and subjective task values interact to explain individuals' achievement and achievement-related choices (e.g., Eccles, 1983; Eccles & Wigfield, 2020; Wigfield & Eccles, 2020).

In addition, these two most proximal motivational factors mediate the influence of other, more distal factors. As can be seen in Figure 1 (in the middle and on the left side), Eccles and colleagues consider various (motivational) factors that indirectly influence individuals' choices and achievement-related performances through expectancies of success and subjective task values. For example, these factors include individuals' socializers (e.g., parents, peers, and teachers), societal aspects (e.g., gender roles and stereotypes) but also achievement-related experiences and gender. Thus, the SEVT offers a comprehensive theoretical framework for studying both proximal psychological processes that operate over short-time frames and the long-term ontogeny of the beliefs and memories underlying individuals' motivated achievement-related choices (Eccles & Wigfield, 2020). Note that the SEVT was initially developed by Eccles and colleagues (1983) to also explain the profound gender differences in achievement-related educational and occupational choices (Eccles & Wigfield, 2020). Given this specific focus and being one of the major guiding theoretical frameworks in this research field, the SEVT was chosen as the main conceptual starting point and theoretical framework for the present doctoral thesis. This was specifically the case in the first research strand in which gender differences in achievement, achievement profiles, and achievement motivation in top-performing math students and their relation to gender equality were examined (Study I).

Figure 1. *Eccles and Colleagues' Situated Expectancy-Value Theory (SEVT) of Achievement Performance and Choice*



Note. Adapted from "35 years of research on students' subjective task values and motivation: A look back and a look forward" by A. Wigfield and J. S. Eccles, 2020, in A. J. Elliot, *Advances in Motivation Science*, Vol. 7, p. 165 (https://doi.org/10.1016/bs.adms.2019.05.002). Copyright 2020 by Elsevier. Reprinted with permission.

However, also in the second research strand, the SEVT was influential in that the functional relation between achievement and one specific aspect of achievement motivation in the SEVT—the academic self-concept—were investigated (Study II).

In sum, the SEVT is the major theoretical framework in my doctoral thesis that helps to explain the emergence of gender differences in achievement and achievement motivation (see Section 1.3.4) and relates these two constructs, particularly with regards to academic self-concepts (see Section 1.4). In the next chapter, I introduce the important topic of gender differences in education and educational trajectories, which presents the first research strand of my doctoral thesis on the interplay of achievement and achievement motivation.

1.3 Gender Differences in Education and Educational Trajectories: The Role of Achievement and Achievement Motivation

In the majority of psychological variables, men and women only differ to a small or negligible degree. However, there are specific domains that show vital differences that can be related to achievement and achievement motivation and have implications for individuals' educational and occupational choices. In this section, I will first provide a terminological classification of the terms *gender* and *sex* and how these terms are used in this doctoral thesis in general (see Section 1.3.1). Then, I will present current research findings on gender differences in education and educational trajectories in the general population (see Section 1.3.2) and in top-performing math students (see Section 1.3.3). Finally, I attempt to elucidate the role of achievement and achievement motivation for gender differences in education and educational trajectories, specifically of the top 5% (see Section 1.3.4) in terms of the SEVT (e.g., Eccles, 1994; Eccles & Wigfield, 2020) and Social Role Theory (e.g., Eagly, 1987; Wood & Eagly, 2012).

1.3.1 Terminology for Comparisons of Men and Women

The terminology for referring to comparisons of men and women is complex and lacks consensus in psychology (Eagly, 2013; Eagly & Wood, 2013). Some researchers have argued that *sex* should be used to refer to female and male biology and *gender* to refer to the sociocultural construction of male and female categories (e.g., Muehlenhard & Peterson, 2011; Unger, 1979; West & Zimmerman, 1987). However, nature and nurture are closely intertwined as causes of differences in the behavior of men and women. Thus, a clear terminological distinction between sociocultural and biological origins is not sensible (e.g., Berenbaum et al., 2011; Eagly & Wood, 2013; Miller & Halpern, 2014). To acknowledge this fact, some researchers have started to use the term *sex/gender* (Schellenberg & Kaiser, 2018) or *gender/sex* (Hyde et al., 2019; van Anders, 2015). Others have tried to end the separation between nature and nurture by using the terms *sex* and *gender* interchangeably (e.g., Hines, 2009; Maccoby, 1988). However, this approach violates the scientific principle of conceptual and terminological accuracy (Glasser & Smith, 2008). Eagly (2013) used the term *sex* in her approach by its common-language meaning of male and female as categories based on a biological distinction, but without causal implications in terms of the nature-nurture debate. The term *gender* is used to refer to the meanings that cultures and individuals *ascribe* to the male and female categories (e.g., gender stereotypes and gender roles). Finally, another approach (Lips, 2008) uses the term *sex* to discuss anatomy and the classification of individuals based on their anatomical category. In this approach, *gender* is used as a more inclusive term for the results of all female–male comparisons, regardless of any causal implications in terms of nature and nurture. Moreover, according to Lips (2008), the term *gender* refers to the societal expectations regarding feminine and masculine roles. In the present doctoral thesis, I adopt

the terminological use of gender *sex* and *gender* by Lips (2008). That is, *gender* is used as an umbrella term for comparisons between men and women and to describe societal expectations regarding female and male categories, whereas the term *sex* is only used for anatomy-related discussions.

1.3.2 Gender Differences in Education and Educational Trajectories in the General Population

In the past few decades, much research has been devoted to the study of gender differences in education and educational trajectories. Whereas early scientists attested women to be intellectually deficient compared to men and motivated mainly by maternal instinct (Shields, 1975), current psychological research suggests that women and men differ in only a few areas. In a seminal synthesis of 46 meta-analyses, Hyde (2005) showed that "males and females are alike on most—but not all—psychological variables" (Hyde, 2005, p. 590), including academic achievement and a range of motivational variables. Of 124 studied effect sizes for gender differences, 78% were small or very close to 0 (*d* between 0 and |0.35|). Based on this finding, Hyde (2005) put forward the *gender similarities hypothesis*, which contradicted popular media reports that emphasized differences between the genders. These findings have been extended and replicated in a more recent synthesis of 106 meta-analyses and 386 effect sizes for gender differences, resulting in 85% of small effect sizes or effect sizes very close to zero (Zell et al., 2015).

Note that meta-analyses are particularly valuable for the estimation of gender differences because they evaluate the magnitude, consistency, replicability, and variability of findings, explore moderators that might contribute to the presence or absence of gender differences, and relate them to relevant psychological theories (e.g., Eagly, 2013; Hyde, 2014). For example, results of a multitude of meta-analyses demonstrated that (in the

general population) differences in the achievement of males and females in mathematics are on average negligible (i.e., *d* between 0 and |0.10|; e.g., Baye & Monseur, 2016; Else-Quest et al., 2010; Hyde, Fennema, & Lamon, 1990, 2008; Lindberg et al., 2010; Reilly et al., 2015, 2019), challenging the stereotype that mathematics is a male domain.

Study I in this doctoral thesis provides a review of gender differences in achievement and achievement motivation in mathematics, reading, and science that were obtained from previous meta-analyses and large-scale studies (see Tables S1 and S2 in Supplemental Online Materials [SOM] of Study I). Table 1 displays the proportions of the reviewed effect sizes for gender differences in achievement and achievement motivation in mathematics, reading, and science that were categorized as negligible, small, moderate, large, or very large according to the benchmarks by Hyde (2005). The table shows that females and males seem to perform equally well in standardized tests in mathematics, reading, and science. The vast majority of gender differences in mathematics, reading, and science achievement in the general population were negligible or small (i.e., mathematics = 100%, reading = 95%, science = 69%). Furthermore, males and females did not differ substantially in their achievement motivation in mathematics, reading, and science. Gender differences in mathematics motivation were in almost all reviewed studies (94% of all effect sizes) small or close to zero. Although about half of the gender differences in reading and science motivation were also negligible or small, 16% of the effect sizes for gender differences in science and 17% of the effect sizes for gender differences in reading fell into the large to very large range. Large to very large differences were found in females' and males' interest in engineering (d = 0.83 to 1.11;[1] Su et al., 2009; Su & Rounds, 2015), in their interest in engineering technology (d = 0.89; Su & Rounds, 2015),

[1] Positive values indicate an advantage of males, negative values an advantage of females.

in their interest in mechanics and electronics (d = 1.21; Su & Rounds, 2015), and in their enjoyment of reading (d = -0.67; OECD, 2010).

Table 1. *Proportion of Effect Sizes (in Percent) for Gender Differences in Achievement and Achievement Motivation in Mathematics, Reading, and Science That Are Negligible, Small, Moderate, Large, or Very Large*

	Achievement			Achievement motivation		
Magnitude	Math[a]	Reading[b]	Science[c]	Math[d]	Reading[e]	Science[f]
Negligible	52	38	23	22	0	35
Small	48	57	46	72	50	27
Moderate	0	5	31	6	33	23
Large	0	0	0	0	17	8
Very large	0	0	0	0	0	8

Note. Figures may not add up to 100% because of rounding.
Negligible = $0.00 < |d| \leq 0.10$, small = $0.10 < |d| \leq 0.35$, moderate = $0.35 < |d| \leq 0.65$, very large = $|d| > 1.00$. k = Number of effect sizes, n = number of studies.
[a] k = 1905, n = 13
[b] k = 1008, n = 12
[c] k = 1264, n = 7
[d] k = 1258, n = 10
[e] k = 201, n = 5
[f] k = 847, n = 7

The review of gender differences in achievement and achievement motivation in Study I also illustrates that the magnitude of the gender gaps varied across countries. There is evidence that these cross-national variations depend on the sociocultural context (for the theoretical foundations, see Section 1.3.4). For example, research shows that gender differences in students' mathematics achievement were smaller in countries with higher levels of gender equality (e.g., Guiso et al., 2008) or a higher share of women in research positions (e.g., Else-Quest et al., 2010). However, other studies found that gender differences were actually more pronounced in more gender equal societies (e.g., Reilly, 2012; Reilly et al., 2019; Stoet & Geary, 2018). Yet, these heterogenous findings might be related to the selection of different gender equality indicators (see also Section 4.2.3) and

varying analytical decisions between studies (see also Section 4.4.2.2 in the General Discussion).

When it comes to school and university education, female students, on average, even appear to be more successful than male students. Indeed, recent research shows that female students have caught up with or even surpassed male students with regard to their performance on school-based assessments (Voyer & Voyer, 2014). Interestingly, studies that investigated gender differences in students' achievement profiles[2] showed that male students demonstrated stronger math achievement tilts than female students; however, female students demonstrated stronger verbal achievement tilts than their male counterparts in school subjects at the age of 16 (Dekhtyar et al., 2018) and in college entrance exams (Coyle et al., 2014, 2015). Overall, there are more women (59%) than men who graduate with a Bachelor's degree across all member states of the European Union (Eurostat, 2020; for similar results in the U.S., see Meece & Askew, 2012).
Importantly, STEM (Science, Technology, Engineering, and Mathematics) is an area in which substantial gender differences in tertiary education and later occupations are observed. STEM is given special attention because STEM professions are important to a country's innovation and prosperity (BMBF, 2019; Halpern et al., 2007; National Science and Technology Council, 2018). In addition, new jobs are created at a faster pace in STEM than jobs in other occupations and wages in STEM are higher than wages in other sectors (Noonan, 2017). Despite its significance, women are still underrepresented in specific STEM domains. This is, for example, reflected in the percentage of male and female university graduates in different academic disciplines. In 2017, fewer women than men

[2] Achievement profiles are composed of the pattern and structure of achievement in several domains within an individual. One way to create achievement profiles is to calculate achievement tilts by subtracting a student's test score in one domain from the same student's test score in another domain (see also Section 1.5.1).

earned Bachelor's degrees in engineering, computer sciences, and physics (20–29% women); on the contrary, more women than men obtained Bachelor's degrees in education (84% women), biology (69% women), social sciences (68% women), biochemistry (65% women), and medicine (61% women) in the EU (Eurostat, 2020). Similar results were reported for the U.S. in 2016 (National Science Foundation, 2019). Notably, the proportion of women graduating with a Bachelor's degree in computer science even decreased from 36% in 1983 to 27% in 1997, and most recently to 19% in 2016 in the U.S. (National Science Board, 2006; National Science Foundation, 2019).

In sum, females and males in the general population perform similarly in mathematics, reading, and science and show similar levels of achievement motivation in mathematics. Even though female students have caught up and to some extent even surpassed male students in school-related performance. Nevertheless, there are notable gender differences in female and male students' achievement profiles and to some extent also in their reading and science motivation. Importantly, female and male students make different educational choices in that, for example, more male students choose physics, engineering, and computer science courses at university level. In the next chapter, I address potential gender differences in top-performing math students.

1.3.3 Gender Differences in Education and Educational Trajectories in Top-Performing Math Students

Mathematics is a prominent content domain that is to some extent still considered as a male domain (e.g., Cvencek et al., 2014), even though the gender gap in mathematics achievement has almost vanished over the years in the general population (e.g., Lindberg et al., 2010). Special attention has been given to gender differences in the group of top-performers in mathematics as they are most likely to enter STEM fields (Park et al., 2007),

and in fact women are still reported to be underrepresented in STEM (Eurostat, 2020; Halpern et al., 2007; National Science Foundation, 2019).

As shown in the review of gender differences in achievement and achievement motivation in Study I (see Tables S1 and S2 in Appendix I), the systematic evidence base was weak for gender differences in top-performing math students. Meta-analyses that covered highly selective samples (Hyde, Fennema, & Lamon, 1990; Lindberg et al., 2010) and students in the top 5% in mathematics (Baye & Monseur, 2016) indicated that gender gaps in mathematics were in those samples somewhat larger ($0.15 \leq d \leq 0.54$) than in the general population ($-0.05 \leq d \leq 0.31$; see Tables S1).

Furthermore, studies show a higher share of male students among top-performing math students in representative samples from the U.S. between 1960 and 1994 (female-to-male ratio of 1:1.50 to 1:4.09; Hedges & Nowell, 1995; Nowell & Hedges, 1998). Studies that used more recent national and international representative samples from 1990 to 2011 found somewhat more balanced female-to-male ratios in the top 5% in mathematics (1:1.09 to 1:2.13; Hyde et al., 2008; Machin & Pekkarinen, 2008; Reilly et al., 2015; Stoet & Geary, 2013). Similarly, talent search studies reported that the preponderance of male students in the highest levels of math achievement (top 0.5%) still exist, but that it rapidly declined from a female-to-male ratio of 1:2.61 in the early 1980s to a ratio of 1:1.37 in the early 2010s (Makel et al., 2016). Another talent search study by Olszewski-Kubilius and Lee (2011) reported slightly higher female-to-male ratios (1:2.5 to 1:3.7) for students in the top 2% in mathematics between 2000 and 2008.

Evidence from meta-analyses or large-scale studies on gender differences in top-performing math students' achievement motivation are scarce (see Table S2 in the SOM of Study I). Only one meta-analysis has investigated gender differences in math anxiety in highly selected samples, indicating that females in this group of students reported a slightly

lower math anxiety than female students in the general population (Hyde, Fennema, Ryan et al., 1990).

Evidence on gender differences in top-performing math students' achievement-related and motivational profiles is only available from single studies or studies using data from talent search programs that are affected by selection effects. For example, the achievement of academically talented male students tilted more toward mathematics than the achievement of their female counterparts; in contrary, the achievement of academically talented female students tilted more toward verbal domains than the achievement of academically talented male students did. These gender differences in achievement tilts seemed to increase with students' achievement level (i.e., from the top 5% to the top 1% to the top 0.01% of ability; Wai et al., 2018). A study by Wang et al. (2013) showed that female students with high math scores tended to also have high verbal scores, whereas male students with high math scores were less likely to have high verbal scores.

In sum, prior studies revealed gender differences in the group of top-performers in mathematics, however, such evidence is weakening with time. There is still little generalizable knowledge about gender differences in top-performing math students' achievement motivation and achievement profiles, specifically using meta-analytic analyses of international large-scale assessments. There is no study yet that has comprehensively and comparatively meta-analyzed gender differences in achievement, achievement motivation, and motivation profiles in this group of students.

1.3.4 How Can We Explain Gender Differences?

The SEVT assumes that gender stereotypes in a society (e.g., that mathematics is a male domain; e.g., Cvencek et al., 2011, 2014; Nosek et al., 2002) influence both directly and indirectly through socializers' beliefs and behaviors, individuals' perception of their

gender roles, characteristics and demands of activities, and possible activities for their own gender (boxes at the top left in Figure 1). More specifically, the SEVT suggests that these perceptions have an impact on individuals' goals and self-schemata, which subsequently influence their expectancies of success (boxes at the top in the middle and right in Figure 1) and subjective tasks values (box at the bottom right in Figure 1). For example, if the gender stereotype that mathematics is a male domain prevails in a society, parents may provide more math-related learning opportunities and experiences to their sons than to their daughters (e.g., Jacobs & Bleeker, 2004). Consequently, a girl learns that mathematics is not consistent with her gender role, which makes her less likely to value, engage, and feel competent in mathematics. This likely results in that the girl performs more poorly in mathematics. Ultimately, this girl is not likely to choose a math-related education or future career, but rather a field that she values and feels competent in (e.g., the humanities). Thus, according to the SEVT, the most important factor to explain gender differences in educational outcomes are not gender comparisons within a domain, but domain comparisons within individuals (Eccles, 1994). Yet, to the extent that female and male students systematically differ in their intraindividual hierarchies of expectancies of success and subjective task values, gender differences in their performance and educational and occupational choices should emerge.

However, while the SEVT can single out the specific gender-specific socialization processes that influence students' achievement and achievement motivation and consequently lead to gender differences in educational contexts, it lacks specification to explain how gender stereotypes—the central factor for gender differences in outcomes according to the SEVT—emerge. Note that the *Social Role Theory* (SRT; e.g., Eagly, 1987; Wood & Eagly, 2012) in combination with the related *Role Congruity Model* (Diekman et al., 2010, 2011) may provide a complementary or alternative explanation for

gender differences. According to the SRT, people infer characteristics of men and women from the roles that men and women typically inhabit in a society. The historical division of labor between men (being the breadwinner) and women (bearing and nursing children) has led to the association of women to be warm, caring, and socially skilled and men to be assertive, dominant, and forceful. Gender role beliefs convey these attributes as generally desirable and admirable for each sex. They encourage children to acquire the skills, characteristics, and preferences that support their society's division of labor through norms and socialization practices. To the extent that people internalize the roles related to their gender, they develop gender identities that let women perceive themselves as especially communal and men as especially agentic. These gender identities and related personal goals are supposed to regulate men's and women's engagement in tasks or occupations that offer opportunities to meet communal or agentic goals (Eagly, 1987; Sczesny et al., 2019; Wood & Eagly, 2012). Specifically based on the related role congruity model (Diekman et al., 2011), women, for example, tend to select and pursue role-congruent, communal goals (i.e., working with or helping others). Applied to the school context, female students should be more attracted to language-related subjects (e.g., native language or foreign language education) than male students because these subjects fulfill communal values (e.g., focus on communication with other people). Moreover, if women's roles in a society do not comprise math-intensive tasks or occupations, we infer from these observations that mathematics, for example, is not for women, which becomes entrenched in gender stereotypes. In general, the SRT predicts that gender differences in a society (e.g., in mathematics achievement) should be smaller in more gender equal societies. As the SEVT states that socialization processes lead to gender differences in male and female students' domain-specific achievement and motivation. As a consequence, socialization processes

that are less gender-typed should produce smaller gender differences in achievement and motivation (for further details, see Study I).

Taken together, the SRT explains the psychological mechanisms that lead to gender stereotypes and how gender-typed roles influence gender differences in educational contexts; the SEVT is more focused on explaining how gender-typed roles influence gender differences in achievement and motivation. Note, however, the SEVT in its most recent edition includes gender and other social role systems as factors of the cultural milieu (see top left box in Figure 4); nonetheless, the specific role of gender for achievement-related performance and choices awaits further elaboration. To this end, it may be reasonable for the SEVT to adopt the hypothesis of the role congruity model (Diekman et al., 2011) that women and men tend to select and pursue goals that are congruent to their gender-specific roles.

1.4 The Relation Between Achievement and Corresponding Academic Self-Concepts

In the present doctoral thesis, I investigate a second research strand on the interplay between achievement and achievement motivation, with the latter including academic self-concepts. Academic self-concepts are defined as a person's mental representations of his or her own abilities in academic domains (Marsh & Craven, 1997). They are considered as one type of achievement motivation and as central mediating constructs in educational-psychological research that affect various psychological and behavioral outcomes. Studies have shown that academic self-concepts are related to students' educational attainment (Guay et al., 2004), course selection (Marsh & Yeung, 1997), interests (Marsh et al., 2005), and learning processes (Byrne, 1996). The SEVT assumes a reciprocal relation between individuals' achievement and their academic self-concepts. The theory suggests

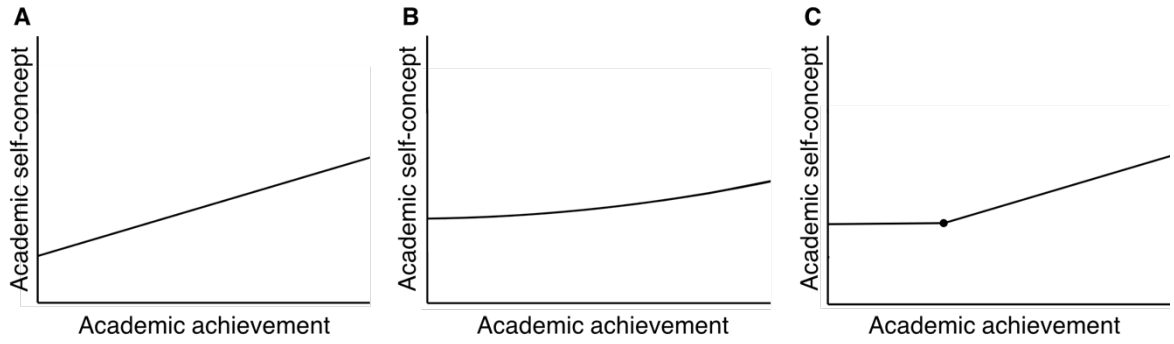
that individuals' achievement (box bottom left in Figure 1, "Previous achievement-related experiences") influences their academic self-concepts (box top in the middle in Figure 1) via their interpretations of their achievement (box bottom in the middle in Figure 1). Academic self-concepts are both influenced by social cognitive processes (i.e., causal attribution processes; *Is my success a consequence of high ability, hard work, luck, or a benevolent evaluation of the teacher?*) and different sources of information. These sources of information include social comparisons (e.g., *How did my classmates do on the assignment?*), dimensional comparisons across domains (e.g., *Do I perform better in mathematics than in English?*), and temporal or internal comparisons across time (e.g., *Have I improved my performance compared with the last exam?*; Eccles & Wigfield, 2020; Wigfield et al., 2020). To make the relation between achievement and corresponding self-concepts reciprocal, the SEVT further proposes that individuals' academic self-concepts influence their future achievement (box on the right in Figure 1) through individuals' expectancies of success (box top right in Figure 1).

Research on academic self-concept formation is highly influenced by the scholarship of Marsh and colleagues. They developed several models and theories that focus on how various sources of information are integrated by individuals to form their academic self-concepts. Among them are the *Internal/External Frame of Reference* (I/E) Model (Marsh, 1986), the *Reciprocal Effects Models* (REM; Marsh & Martin, 2011), the *Dimensional Comparison Theory* (Möller & Marsh, 2013), and the *Big Fish Little Pond Effect (BFLPE) Model* (Marsh, 1987). Currently, efforts are being made to integrate the research lines around SEVT and self-concept formation (Eccles & Wigfield, 2020; Wigfield et al., 2020).

Although research with children and adolescents has shown that academic achievement has a strong impact on students' academic self-concepts (Möller et al., 2009;

2020), the SEVT (or any other model of self-concept formation) has not fully explicated the functional relation between students' achievement and their corresponding academic self-concepts. Does the functional relation between achievement and academic self-concepts matter? Figure 2 illustrates that different functional relations between achievement and corresponding self-concepts lead to fundamentally different predictions.

Figure 2. *Hypothetical Plots Show the Functional Relation Between Achievement and Corresponding Academic Self-Concepts in Different Statistical Models*



Note. Panel A shows a linear relation between achievement and self-concept, Panel B shows a quadratic relation, and Panel C shows an interrupted relation.

In the *linear model* (Figure 2A), a constant amount of increase in achievement is associated with a constant increase in the corresponding self-concept across the entire achievement continuum. The linear model would predict that the better students perform in a domain, the higher their corresponding academic self-concept. Consequently, the worst performing student in a class would report the lowest academic self-concept and the best performing student the highest academic self-concept. This functional relation would be plausible if students solely used social comparisons to draw conclusions on their abilities (Festinger, 1954; Gerber et al., 2018).

In the *quadratic model* (Figure 2B), students' academic self-concept does not (or barely) change up to a certain point, even though students' performance increases; thereafter, students' academic self-concept increases as their achievement increases. Thus, increments in achievement have different effects on students' academic self-concepts depending on how well or badly they perform. The quadratic model would predict that the increase in students' self-concept is weaker for lower achieving students than for higher achieving students.

In an *interrupted regression model* (Figure 2C), the data is divided into k bins and for each of the k bins a category-specific linear model is simultaneously specified. For example, Figure 2C shows $k = 2$ bins. These linear models are continuous at $k - 1$ joint-points, called knots. Hence, for different sections on the achievement continuum, different amounts of increase in achievement are associated with a different increase in the corresponding self-concept. For example, the interrupted regression model shown in Figure 2C would make the following prediction: Students' achievement is not related to their self-concepts for lower achieving students, but the relation is positive for higher achieving students.

One major reason for why the relations between achievement and self-concepts might vary as a function of individual student achievement (Figures 2B and 2C) is that being asked to evaluate one's own abilities in self-concept questionnaires may trigger self-protective strategies. Intuitively, we would expect that students with lower achievement in a specific domain should have lower evaluations of their abilities in this domain. However, a negative self-evaluation is a major threat to the self. To protect their self-worth, individuals with low achievement are likely to engage in self-protective strategies that result in more positive self-views (Alicke & Sedikides, 2009).

However, in a review presented in Study II, we showed that in all studies only linear relations between achievement and self-concept were analyzed and nonlinear relations were not considered (see Study II). It is important to examine also nonlinear relations between the constructs because linear models might not fully capture the relation between achievement and the corresponding self-concept for a large part of the student body. To this end, Study II aimed to systematically examine the form of the functional relation between achievement and corresponding self-concepts. As I will argue, the revealed functional form of the relation has implications for the assessment and interpretation of self-concepts and ultimately for models of self-concept formation.

1.5 Current Methodological Approaches to Study the Interplay Between Achievement and Achievement Motivation

In the previous Sections 1.3 and 1.4, I have outlined the two research strands of Studies I and II on examining the relationship between achievement and achievement motivation and relevant prior research. In this section, I introduce two current methodological approaches in psychological research that are relevant for the doctoral thesis: First, approaches are presented that enable us to model intraindividual hierarchies of achievement and achievement motivation; second, systematic research syntheses with individual participant data are presented that allow us to estimate the robustness and generalizability of results across studies.

In this doctoral thesis, I combined both methodological approaches to study the (intraindividual) interplay between achievement and achievement motivation across samples and studies.

1.5.1 Modelling Intraindividual Hierarchies of Achievement and Achievement Motivation

In the SEVT, it is assumed that intraindividual hierarchies in achievement and achievement motivation play an important role for individuals' subsequent achievement and achievement-related choices (Eccles, 1994; Wigfield & Eccles, 2020). There are several approaches to assess these intraindividual hierarchies and to examine their relation to different educational outcomes statistically. The approaches can be grouped in person-centered and variable-centered approaches as well as less and more complex approaches. In the following, I will first present person-centered approaches and subsequently variable-centered approaches to model intraindividual hierarchies of achievement and achievement motivation.

Person-centered approaches take into account that one sample may consist of several subpopulations for which different sets of estimates can be computed (Morin et al., 2016). Thereby, both the actual and relative levels of one variable to another are considered to form homogeneous groups (Meece & Agger, 2018). Person-centered analyses can be performed with mixture models, including cluster analysis, latent profile analyses, latent class analyses, latent transition analyses, mixture regression, and growth mixture models—statistical analyses that have in common that they attempt to detect subgroups of individuals who are similar in a number of characteristics. Note that person-centered approaches are increasingly used to model the complex interplay between abilities and achievement motivation (e.g., Conley, 2012; Lazarides et al., 2020; Musu-Gillette et al., 2015; Wang et al., 2013; Watt et al., 2019; but see also Bauer & Curran, 2003). For example, Wang et al. (2013) performed a latent profile analysis of math and verbal scores

and identified five 12th-grade competence profiles (e.g., a high-math/high-verbal profile) to predict the choice of a STEM occupation in each profile.

In contrast, *variable-centered approaches* assume that all individuals from a sample originate from a single population for which averaged estimates can be calculated (Morin et al., 2016). A variable-centered approach that can capture the complex interplay between achievement and achievement motivation is the integrative trait-complex approach (Ackerman et al., 2013). Here trait complexes are identified in an exploratory factor analysis. As a first step, the underlying factors for a set of measured scales are determined. Then, unit-weighted *z*-score composites are calculated from the scales that have salient loading on the respective factors (e.g., Ackerman, 2003; Ackerman et al., 2013). These scores can be considered as profile scores. Thus, the integrative trait-complex approach can both model the interplay between achievement and achievement motivation as well as the intraindividual hierarchies between these trait complexes. For example, Ackerman et al. (2013) identified 5 trait complexes from 29 scales of which the first trait complex "math/science self-concept" included the scales math self-concept, spatial self-concept, science self-concept, self-estimates of math ability, and numerical preferences.

Less complex approaches to assess intraindividual hierarchies in individuals' achievement (i.e., achievement profiles) are, for example, the division into achievement groups (e.g., Lubinski et al., 2001) or the calculation of relative academic strengths (Coyle et al., 2014, 2015; Dekhtyar et al., 2018; Park et al., 2007; Stoet & Geary, 2018). To form achievement groups, Lubinski et al. (2001), for example, divided academically talented students into three different achievement groups based on their SAT profiles. High-verbal students had verbal SAT scores that exceeded their math SAT scores by more than one standard deviation, high-math students had math SAT scores that exceeded their verbal SAT scores by more than one standard deviation, and high-flat students had math and

verbal SAT scores that fell within one standard deviation of each other. Studies that analyzed relative academic strengths differ with regard to how the within-subject difference between different test scores is calculated (number of academic domains, standardized difference or not), and how balanced their achievement profiles are (e.g., equal or inequal numbers of numeric/science/technical and verbal domains; Coyle et al., 2014, 2015; Dekhtyar et al., 2018; Park et al., 2007; Stoet & Geary, 2018).

Another simple option to assess intraindividual hierarchies in individuals' achievement motivation is to use "comparative" survey questions (also called forced-choice formats) that ask individuals, for example, whether they value math or English more (e.g., by applying rank-ordering items). However, these items have so far scarcely been used to measure intraindividual hierarchies in students' achievement motivation (Wigfield & Eccles, 2020).

To conclude, there is a variety of person- and variable-centered approaches that can capture the intraindividual hierarchies in students' achievement and achievement motivation. Depending on the level of (and complexity of) the research question, there is quite a wide range of assessment options available. In Study I, we chose a variable-centered approach and created three achievement profiles (i.e., math–reading, science–reading, math–science) by calculating students' relative academic strengths in two different academic domains.

1.5.2 Research Synthesis with Individual Participant Data

A second recent development in psychological research is to consider the robustness and generalizability of research results in research syntheses. This is important as study results usually vary to different degrees. Scientific phenomena are typically examined in multiple studies to investigate their replicability and external validity (i.e., *Do effects hold over*

variations in persons, settings, treatments, and outcomes?; Cooper et al., 2019; Shadish et al., 2002). Of note, research results are rarely identical even in direct replication attempts with a high level of precision (Open Science Collaboration, 2015; Valentine et al., 2011) both in behavioral sciences and "harder" sciences such as physics (Hedges, 2019). The variation in study results (or the failure to replicate results) may be explained by differences in statistical power, sampling variation, measurement error, insufficient construct validity, and others (Cooper et al., 2019; Schmidt & Oh, 2016; Shrout & Rodgers, 2018). Yet, variation in study results becomes problematic if the variation is due to scientific misconduct (e.g., selective reporting of results, *p*-hacking; Simmons et al., 2011). This lack of replicability through questionable research practices led to a proclamation of the so-called "replication crisis" and a public loss of confidence in many scientific fields, including psychology (Ioannidis, 2005; Pashler & Wagenmakers, 2012).

Systematic research synthesis is one critical tool to investigate cross-study variation (or heterogeneity; Roisman & van IJzendoorn, 2018; Schmidt & Oh, 2016; Shrout & Rodgers, 2018; but see also Nelson et al., 2018). More specifically, research syntheses summarize the results of scientific (replication) studies to make generalizations, but also to explore the boundary conditions of generalizations (Cooper et al., 2019). Using meta-analytical techniques, weighted average effect sizes, variations in effect sizes between studies, and factors that moderate the size of the effects can be examined and estimated.

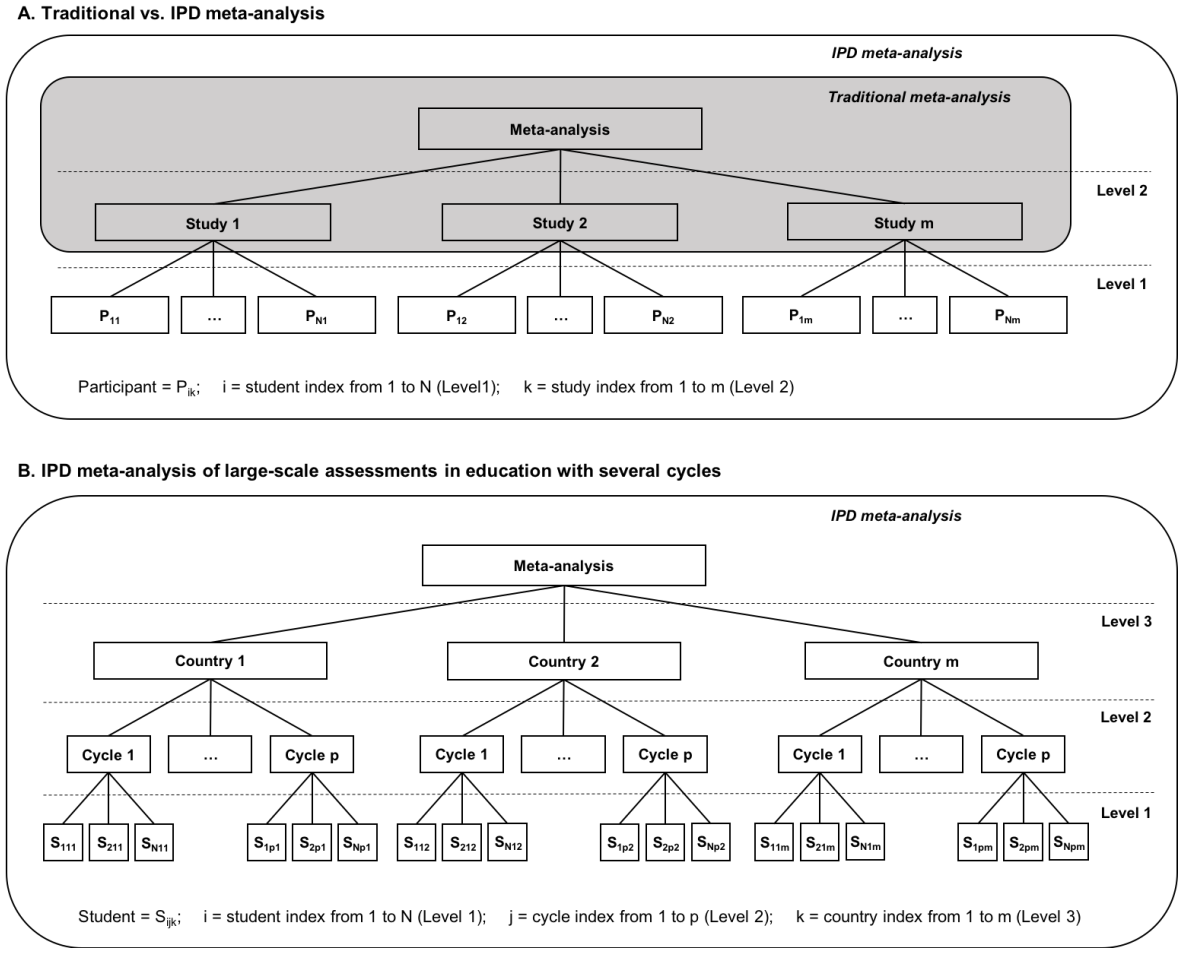
The use of quantitative techniques to integrate empirical studies dates back to the early 18th century, when the English mathematician Roger Cotes computed weighted averages of astronomical measurements that were assessed by different astronomers. Since then, such methods were rarely used until Glass (1976) introduced effect sizes as a common metric over studies and coined the term *meta-analysis* (Shadish et al., 2002). In traditional meta-analyses, data are synthesized on an aggregate study level obtained from

study publications or study authors (e.g., an effect size and a standard error or confidence interval; illustrated in grey in Figure 3A).

More recently, the use of meta-analysis has been extended to individual participant data (IPD) synthesis (L. A. Stewart & Parmar, 1993), also called integrative data analysis (IDA; Curran & Hussong, 2009). Compared with traditional meta-analyses, IPD meta-analysis adds data on the participant level to the analyses (Figure 3A; Riley et al., 2010). Thus, IPD meta-analysis involves obtaining and then synthesizing raw data for the individual participants.

Although IPD meta-analysis has been described as a gold-standard method of meta-analysis for quite some time in the biomedical sciences (L. A. Stewart & Tierney, 2002), it has only recently entered the field of psychology (Roisman & van IJzendoorn, 2018). IPD meta-analysis has several advantages over traditional meta-analyses. First, method heterogeneity between studies—a major biasing factor—is drastically reduced by applying the same inclusion and exclusion criteria across studies and synthesizing the data according to a standardized analysis protocol. By capitalizing on IPD in meta-analyses, it is also possible to disentangle subject-level and study-level sources of heterogeneity in effects (Lyman & Kuderer, 2005; van Walraven, 2010). Furthermore, the number of analytic options beyond the focal effect size under study is much larger in IPD meta-analysis because of the access to the raw data and thus more appropriate or advanced methods can be applied when necessary (Cooper & Patall, 2009; Reily et al., 2010). Access to IPD may also help to improve data quality by accounting for missing data at the individual level (Pigott, 2019) as well as by calculating and incorporating results from

Figure 3. *Schematic Overview of Traditional Meta-Analysis (Grey) Versus Individual Participant Data (IPD) Meta-Analysis (Transparent; Panel A). Panel B Displays the Design of the IPD Meta-Analysis/Integrative Data Analysis Used in Studies I and II in the Present Doctoral Thesis*



unpublished studies, which reduces publication bias (Reily et al., 2010). Finally, another potential benefit of IPD meta-analysis lies in statistical power. Modeling effects over participants instead of over studies potentially increases the power of moderator analyses (Cooper & Patall, 2009; Reily et al., 2010).

IPD meta-analyses can be further divided into one-stage and two-stage IPD meta-analyses. A one-stage IPD meta-analysis combines all IPD and analyzes them in a single step by using a multilevel structure to account for the variation within studies. In a two-stage IPD meta-analysis, analyses are first performed for each study and then combined in

a second step using (multilevel) meta-analytical methods. Results from one-stage and two-stage IPD meta-analyses often do not differ, however, one-stage IPD meta-analysis has the advantage of even greater flexibility and higher statistical power when sample sizes of single studies are small (G. B. Stewart et al., 2012).

For IPD meta-analyses (or IDA) that draw on data from international large-scale assessments in education, a two-stage approach is required to account for methodological characteristics of each assessment cycle (e.g., different weights, numbers of plausible values, numbers of jackknife sampling zones). For instance, in this doctoral thesis, effect sizes are first calculated for each country in each assessment cycle separately. In a second step, the effect sizes are aggregated using multilevel meta-analytical methods (see Figure 3B). Thus, it is possible to disentangle within-country level and between-country level sources of heterogeneity in effects.

In sum, research synthesis allows us to systematically estimate the magnitude of the effect sizes and explain sources of cross-study variation. To this end, it fosters reproducible, rigorous, and transparent research (McNutt, 2014), and can in that regard be considered as one critical tool among others to address the issue of replicability and advance psychology as a field (Roisman & van IJzendoorn, 2018).

1.6 Objectives of the Present Doctoral Thesis

In this doctoral thesis, I aimed to investigate the interplay between achievement and achievement motivation. To address this topic, I have chosen two important strands of research that provide different angles on the interplay between achievement and achievement motivation within the framework of the SEVT. The presented research questions were examined by capitalizing on data from international large-scale assessments.

The *first strand of research*, which I covered in Study I, was the extent to which gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation exist. While prior studies have largely examined single, isolated domains and have had a strong focus on U.S. samples, a systematic, meta-analytical analysis of these gender differences in the group of top-performing math students across countries is lacking. To this end, I aimed to tackle in Study I the following research question:

Research Question 1: *What is the extent of gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in mathematics, reading, and science across countries?*

By applying a two-stage multilevel random-effects IPD meta-analysis of representative individual student data, the main goal was to provide reliable and widely generalizable empirical knowledge about the direction, size, and variability of these gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in these three core academic domains for a large number of countries. To this end, these analyses were conducted in the group of top-performing math students (top 5%) by drawing on six cycles from PISA (2000–2015, N = 115,481, 15-year-olds, 82 countries).

Furthermore, it is unclear how gender differences in top-performing math students emerge. One potential reason for cross-national variability in gender differences in this group of students are varying sociocultural factors, such as the level of gender equality in a country. SEVT and SRT predict that gender differences should be smaller in more gender equal societies than in less gender equal societies. To this end, I tackled in Study I a second research question:

Research Question 2: *To what extent are cross-national gender differences in the group of top-performing math students related to sociocultural factors, or more specifically, to the level of gender equality in a country?*

Using the same data as for Research Question 1, the goal was to examine the moderating role of different gender equality indicators for gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation. Gender differences were synthesized in a two-stage multilevel mixed-effects IDP meta-analyses that examined the moderating effects of nation-level gender equality indicators in multivariate meta-regressions. Domain-specific gender equality indicators (i.e., primary, secondary, tertiary enrollment ratios and women's share of higher positions and research positions) were selected from the ILO, the OECD, the UN, and UNESCO.

The *second research strand*, which I tackled in Study II, refers to the question of how achievement and academic self-concept—a central motivational construct in educational psychology—are functionally related. The relationship between achievement and corresponding self-concepts is a critical aspect of the SEVT, but also of other prominent theories of self-concepts formation. Researchers implicitly assume the relation between achievement and corresponding self-concepts to be linear. Although assuming a nonlinear relation between achievement and corresponding self-concepts is highly plausible because of individuals' use of self-protective strategies in self-evaluative situations, the functional relation between these constructs has not yet been systematically examined. To this end, I aimed to tackle the following research question in Study II:

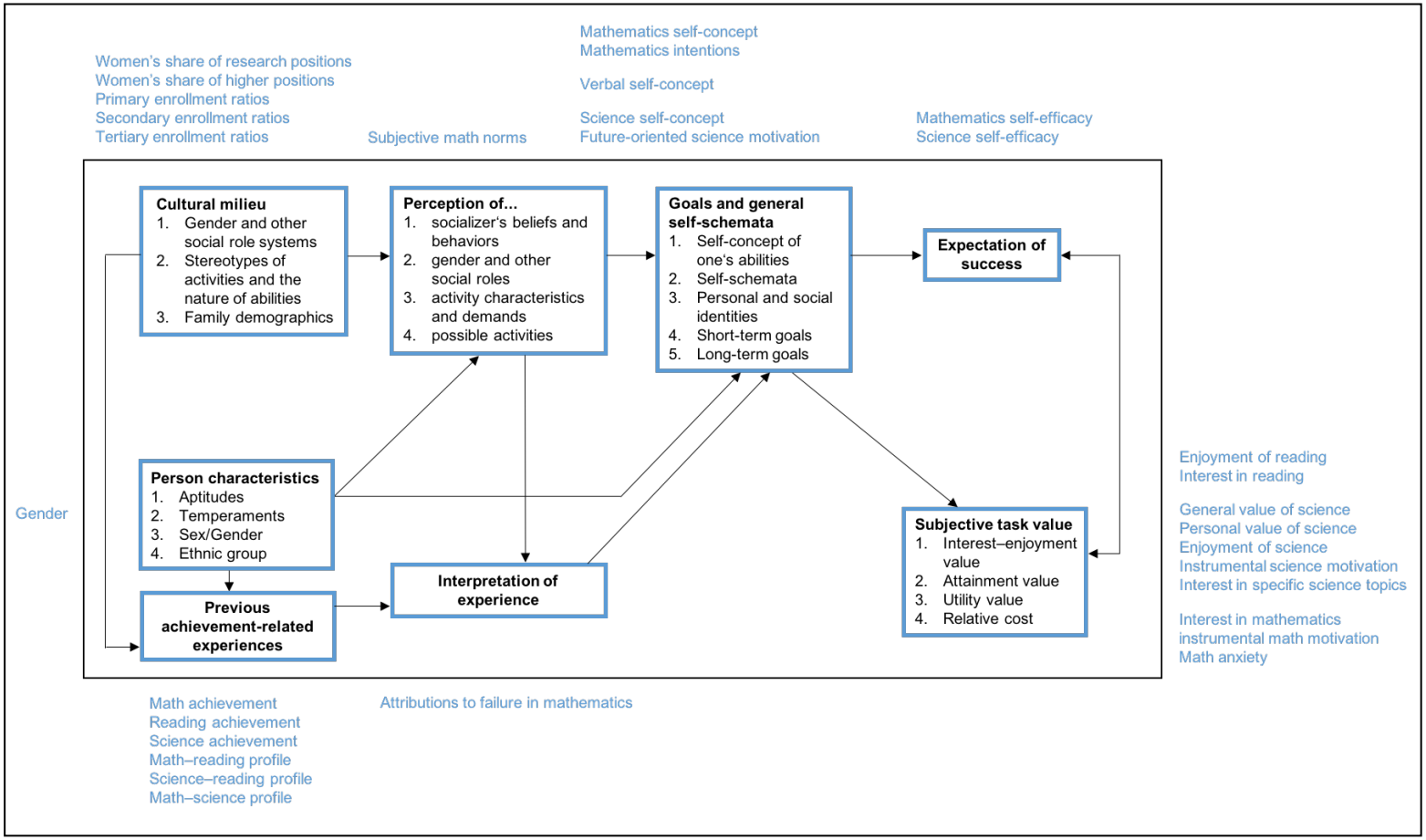
Research Question 3: *Which functional relation exists between students' academic achievement and corresponding academic self-concepts?*

The goal was to examine whether relations between achievement and corresponding self-concepts are nonlinear and to what extent the nonlinearity is

generalizable across different domains, age groups, countries, and analytical approaches in an integrative data analysis. An integrative data analysis investigates the robustness of results by applying the same analysis protocol to several data sets (here: eight cycles from PISA [2000 mathematics and verbal domain, 2003, 2012], TIMSS [2011, 2015], and PIRLS [2011, 2016]). The results were then synthesized in multilevel meta-analytic models. To further examine the generalizability of the results, the functional relation between achievement and corresponding self-concepts was analyzed across two domains (mathematics and the verbal domain), two age groups (elementary and secondary school students) across 13 countries using two analytical approaches (quadratic and interrupted regressions).

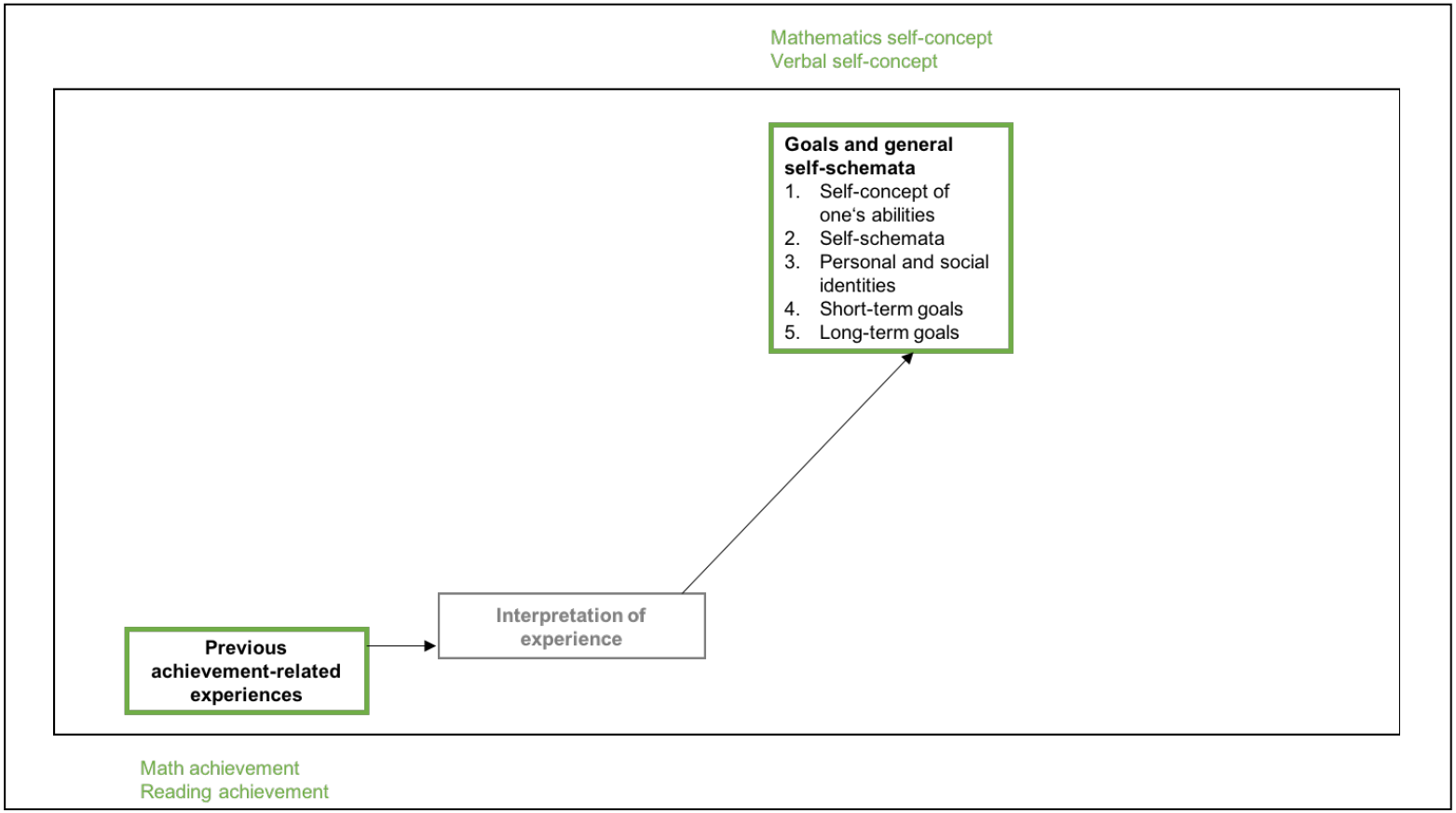
Together, by combining both research strands and answering the above-mentioned research questions, this doctoral thesis aims to foster the understanding of the interplay between achievement and achievement motivation in new ways and to inform the SEVT. To do so, I applied state-of-the-art research synthesis methods on representative high-quality student data, and adopted measures of intraindividual hierarchies in top-performing math students' achievement. To provide an overview of the examined relationships assumed by the SEVT, I illustrated and color-coded them for each study in Figure 4 (Study I) and Figure 5 (Study II). In the figures, only those components of the SEVT are depicted that were covered in the respective studies (for the full model, see Figure 1). As shown in Figures 4 and 5, Study I covered a broad range of gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation related to the SEVT, whereas Study II focused more specifically on two components of the SEVT—achievement and academic self-concepts. Figure 6 provides a combined overview of the examined relationships assumed by the SEVT for the doctoral thesis as a whole.

Figure 4. *Overview of Components of the Situated Expectancy–Value Theory (SEVT) That Were Examined in Study I of the Present Doctoral Thesis.*



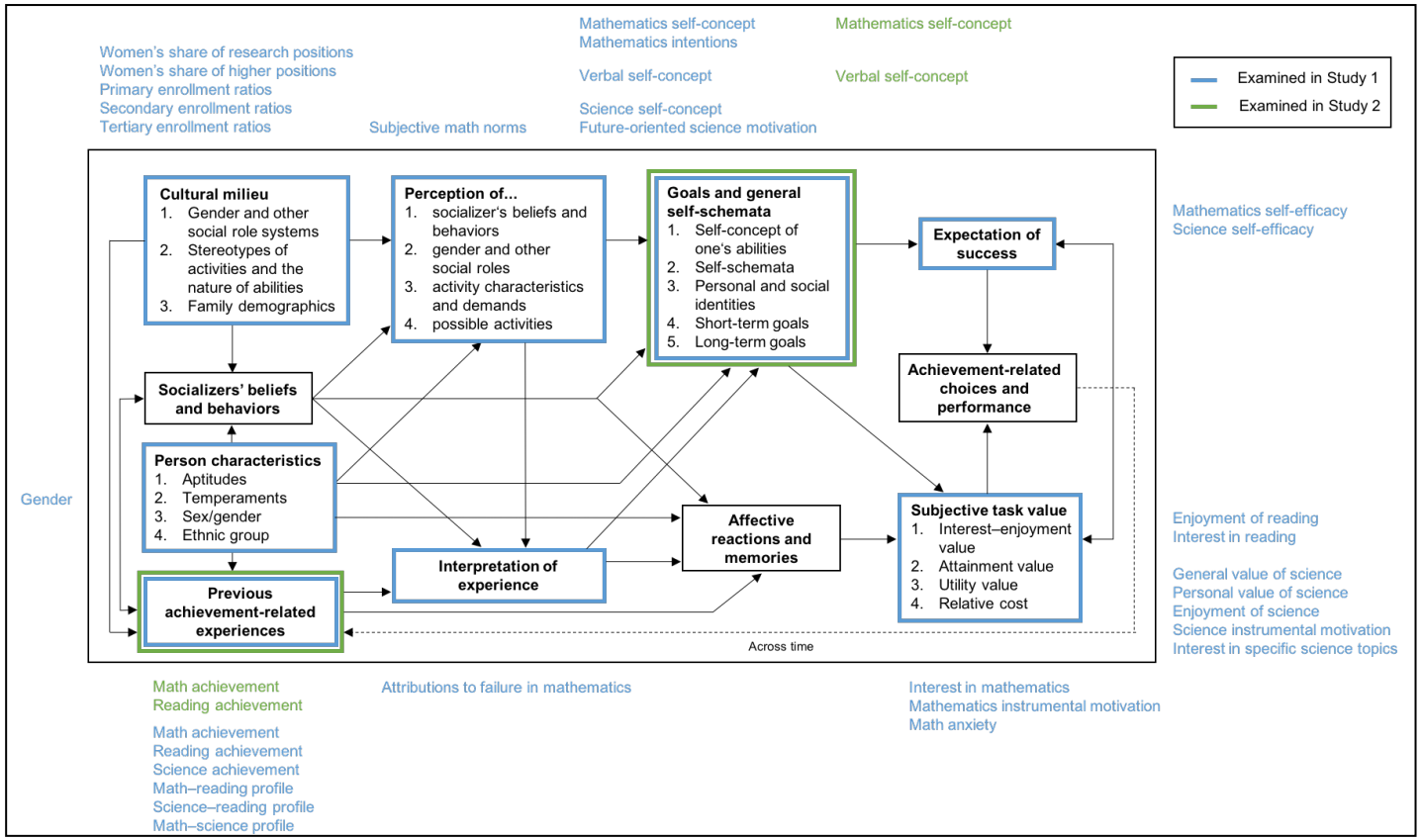
Note. Inner rectangle = Constructs assumed by the SEVT; Outer rectangle = Selected set of variables included in the analysis. Adapted from "35 years of research on students' subjective task values and motivation: A look back and a look forward" by A. Wigfield and J. S. Eccles, 2020, in A. J. Elliot, *Advances in Motivation Science*, Vol. 7, p. 165 (https://doi.org/10.1016/bs.adms.2019.05.002). Copyright 2020 by Elsevier. Reprinted with permission.

Figure 5. O*verview of Components of the Situated Expectancy–Value Theory (SEVT) That Were Examined in Study II of the Present Doctoral Thesis.*



Note. Inner rectangle = Constructs assumed by the SEVT; Outer rectangle = Selected set of variables included in the analysis. Adapted from "35 years of research on students' subjective task values and motivation: A look back and a look forward" by A. Wigfield and J. S. Eccles, 2020, in A. J. Elliot, *Advances in Motivation Science*, Vol. 7, p. 165 (https://doi.org/10.1016/bs.adms.2019.05.002). Copyright 2020 by Elsevier. Reprinted with permission.

Figure 6. *Overview of Components of the Situated Expectancy–Value Theory (SEVT) of Achievement Performance and Choice That Were Studied in the Present Doctoral Thesis. The Inner Rectangle Represents the Constructs Assumed by SEVT, While the Outer Rectangle Represents the Selected Set of Constructs Included in Study I (Blue) and in Study II (Green)*



Note. Adapted from "35 years of research on students' subjective task values and motivation: A look back and a look forward" by A. Wigfield and J. S. Eccles, 2020, in A. J. Elliot, *Advances in Motivation Science*, Vol. 7, p. 165 (https://doi.org/10.1016/bs.adms.2019.05.002). Copyright 2020 by Elsevier. Reprinted with permission.

1.7 References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*(2), 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist, 38*(2), 85–93. https://doi.org/10.1207/S15326985EP3802_3

Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013). Trait complex, cognitive ability, and domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences. *Journal of Educational Psychology, 105*(3), 911–927. https://doi.org/10.1037/a0032338

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66*(2), 91–103. https://www.jstor.org/stable/2112795

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology, 34*(6), 1373–1399. https://doi.org/10.1037/0012-1649.34.6.1373

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*(3), 165–176. https://doi.org/10.1016/j.edurev.2009.04.002

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education, 4*(1). https://doi.org/10.1186/s40536-015-0015

Berenbaum, S. A., Blakemore, J. E. O. & Beltz, A. M. (2011). A role for biology in gender-related behavior. *Sex Roles, 64*, 804–825. https://doi.org/10.1007/s11199-011-9990-8

Bertling, J. P., Borgonovi, F., & Almonte, D. E. (2016) Psychosocial skills in large-scale assessments: Trends, challenges, and policy implications. In A. Lipnevich, F. Preckel, & R. Roberts (Eds.), *Psychosocial skills and school systems in the 21st century.* Springer. https://doi.org/10.1007/978-3-319-28606-8_14

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

Bundesministerium für Bildung und Forschung (2019). *Mit MINT in die Zukunft! Der MINT-Aktionsplan des BMBF.* https://www.bmbf.de/upload_filestore/pub/MINT_Aktionsplan.pdf

Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation.* American Psychological Association.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104*(1), 32–47. https://doi.org/10.1037/a0026042

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610448864.4

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176. https://doi.org/10.1037/a0015565

Coyle, T. R., Purcell, J. M., Snyder, A. C., & Richmond, M. C. (2014). Ability tilt on the SAT and ACT predicts specific abilities and college majors. *Intelligence, 46,* 18–24. https://doi.org/10.1016/j.intell.2014.04.008

Coyle, T. R., Snyder, A. C., & Richmond, M. C. (2015). Sex differences in ability tilt: Support for investment theory. *Intelligence, 50,* 209–220. https://doi.org/10.1016/j.intell.2015.04.012

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Cvencek, D., Meltzoff, A. N., & Kapur, M. (2014). Cognitive consistency and math–gender stereotypes in Singaporean children. *Journal of Experimental Child Psychology, 117*, 73–91. https://doi.org/10.1016/j.jecp.2013.07.018

Dawis, R.V., & Lofquist, L.H. (1984). *A psychological theory of work adjustment: An individual differences model and its application.* University of Minnesota Press.

Dekhtyar, S., Weber, D., Helgertz, J., & Herlitz, A. (2018). Sex differences in academic strengths contribute to gender segregation in education and occupation: A longitudinal examination of 167,776 individuals. *Intelligence, 67,* 84–92. https://doi.org/10.1016/j.intell.2017.11.007

Diekman, A. B., Brown, E., Johnston, A., & Clark, E. (2010). Seeking congruity between goals and roles: A new look at why women opt out of STEM careers. *Psychological Science, 21*(8), 1051–1057. https://doi.org/10.1177/0956797610377342

Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., & Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to stem careers: Evidence for a goal congruity perspective. *Journal of Personality and Social Psychology, 101*(5), 902–918. https://doi.org/10.1037/a0025199

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Erlbaum.

Eagly, A. H. (2013). The science and politics of comparing women and men: A reconsideration. In M. K. Ryan & N. R. Branscombe, *The SAGE handbook of gender and psychology* (pp. 11-28). SAGE Publications Inc. https://doi.org/10.4135/9781446269930.n2

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science, 8*(3), 340–357. https://doi.org/10.1177/1745691613484767

Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of Women Quarterly, 18*(4), 585–609. https://doi.org/10.1111/j.1471-6402.1994.tb01049.x

Eccles (Parsons), 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. 75–146). W. H. Freeman.

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*. https://doi.org/10.1016/j.cedpsych.2020.101859

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Eurostat (2020, May 22). *Graduates by education level, programme orientation, sex and field of education.* https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=educ_uoe_grad02&lang=en

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Fulmer, S. M., & Frijters, J. C. (2009). A review of self-report and alternative approaches in the measurement of student motivation. *Educational Psychology Review, 21*(3), 219–246. https://doi.org/10.1007/s10648-009-9107-x

Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin, 144*(2), 177–197. https://doi.org/10.1037/bul0000127

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.2307/1174772

Glasser, H. M., & Smith, J. P. III. (2008). On the vague meaning of "gender" in education research: The problem, its sources, and recommendations for practice. *Educational Researcher, 37*(6), 343–350. https://doi.org/10.3102/0013189X08323718

Guay, F., Larose, S., & Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity, 3,* 53–68. https://doi.org/10.1080/13576500342000040

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hedges, L. V. (2019). The statistics of replication. *Methodology*, *15*(Supplement 1), 3–14. https://doi.org/10.1027/1614-2241/a000173

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45. https://doi.org/10.1126/science.7604277

Hines, M. (2009). Gonadal hormones and sexual differentiation of human brain and behavior. In D. W. Pfaff, A. P. Arnold, A. M. Etgen, S. E. Fahrbach, & R. T. Rubin (Eds.), *Hormones, brain, and behavior* (Vol. 3, 2nd ed.; pp. 1869–1909). Academic Press.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592. https://dx.doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist, 74*(2), 171–193. https://doi.org/10.1037/amp0000307

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

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14*(3), 299–324. https://doi.org/10.1111/j.1471-6402.1990.tb00022.x

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

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobs, J. E., & Bleeker, M. M. (2004). Girls' and boys' developing interests in math and science: Do parents matter? *New Directions for Child and Adolescent Development, 106*, 5–21. https://doi.org/10.1002/cd.113

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2012). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). Springer. https://doi.org/10.1007/978-94-007-4629-9

Lazarides, R., Dicke, A.-L., Rubach, C., & Eccles, J. S. (2020). Profiles of motivational beliefs in math: Exploring their development, relations to student-perceived classroom characteristics, and impact on future career aspirations and choices. *Journal of Educational Psychology, 112*(1), 70–92. https://doi.org/10.1037/edu0000368

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*(2), 602–640. https://doi.org/10.3102/0034654315617832

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lips, H. M. (2008). *Sex and gender: An introduction* (6th ed.). McGraw–Hill.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*(4), 718–729. https://doi.org/10.1037/0021-9010.86.4.718

Lyman, G. H., & Kuderer, N. M. (2005). The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology, 5*(14). https://doi.org/10.1186/1471-2288-5-14

Maccoby, E. E. (1988) Gender as a social category. *Developmental Psychology, 24*(6), 755–765. https://doi.org/10.1037/0012-1649.24.6.755

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 322*(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Makel, M. C., Wai, J., Peairs, K., & Putallaz, M. (2016). Sex differences in the right tail of cognitive abilities: An update and cross cultural extension. *Intelligence, 59*, 8–15. https://doi.org/10.1016/j.intell.2016.09.003

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/10.2307/1163048

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., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 131–198). Academic Press.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59–77. https://doi.org/10.1348/000709910X503501

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*, 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

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/10.2307/1163354

McNutt, M. (2014). Journals unite for reproducibility. *Science, 346*(6210), 679. https://doi.org/10.1126/science.aaa1724

Meece, J., & Agger, C. (2018). Achievement motivation in education. In *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.7

Meece, J. L., & Askew, K. J. S. (2012). Gender, motivation, and educational attainment. In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences and cultural and contextual factors* (p. 139–162). American Psychological Association. https://doi.org/10.1037/13274-006

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences, 18*(1), 37–45. https://dx.doi.org/10.1016/j.tics.2013.10.011

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., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Morin, A. J., Gagne, M., & Bujacz, A. (2016). Feature topic: Person-centered methodologies in the organizational sciences. *Organizational Research Methods, 19*(1), 8–9. https://doi.org/10.1177/1094428115617592

Muehlenhard, C. L., & Peterson, Z. D. (2011). Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex Roles, 64*, 791–803. https://doi.org/10.1007/s11199-011-9932-5

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

National Science and Technology Council (2018). *Charting a course for success: America's strategy for STEM education.* https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf

National Science Board (2016). *Science and Engineering Indicators 2006 (Vol. 1)*. National Science Foundation.

National Science Foundation (2019). *Women, minorities, and persons with disabilities in science and engineering.* National Science Foundation.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Noonan, R. (2017). *STEM Jobs: 2017 Update* (ESA Issue Brief No. 02-17). U.S. Department of Commerce. https://www.commerce.gov/sites/default/files/migrated/reports/stem-jobs-2017-update.pdf

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83*(1), 44–59. https://doi.org/10.1037//0022-3514.83.1.44

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39*(1–2), 21–43. https://doi.org/10.1023/A:1018873615316

OECD (2010). *PISA 2009 Results: Learning to learn – Student engagement, strategies and practices* (Vol. III). OECD Publishing. https://dx.doi.org/10.1787/9789264083943-en

OECD (2015). *Skills for Social Progress.* OECD Publishing. http://www.oecd-ilibrary.org/content/book/9789264226159-en

Olszewski-Kubilius, P., & Lee, S. Y. (2011). Gender and other group differences in performance on off-level tests: Changes in the 21st century. *Gifted Child Quarterly, 55*(1), 54–73. https://doi.org/10.1177/0016986210382574

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the arts and sciences. *Psychological Science, 18*(11), 948–952. https://doi.org/10.1111/j.1467-9280.2007.02007.x

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pigott, T. D. (2019). Missing data in meta-analysis (pp. 367–382). In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667–686. https://doi.org/10.1037/0022-0663.95.4.667

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS ONE, 7*(7), e39904. https://doi.org/10.1371/journal.pone.0039904

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

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Investigating gender differences in mathematics and science: Results from the 2011 Trends in Mathematics and Science Survey. *Research in Science Education, 49*(1), 25–50. https://doi.org/10.1007/s11165-017-9630-6

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *Bmj, 340*, c221. https://doi.org/10.1136/bmj.c221

Roisman, G. I., & van IJzendoorn, M. H. (2018). Meta-analysis and individual participant data synthesis in child development: Introduction to the special section. *Child Development, 89*(6), 1939–1942. https://doi.org/10.1111/cdev.13127

Schellenberg, D., & Kaiser, A. (2018). The sex/gender distinction: Beyond f and m. In C. B. Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA handbook of the psychology of women: History, theory, and battlegrounds* (p. 165–187). American Psychological Association. https://doi.org/10.1037/0000059-009

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4*(1), 32–37. https://dx.doi.org/10.1037/arc0000029

Schunk, D. H., & Mullen, C. A. (2013). Motivation. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 67–69). Routledge.

Sczesny, S., Nater, C., & Eagly, A. H. (2019). Agency and communion: Their implications for gender stereotypes and gender identities. In A. E. Abele & B. Wojciszke (Eds.) *Agency and communion in social psychology. Current issues in social psychology* (pp. 103–116). Routledge.

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

Shields, S. A. (1975). Functionalism, Darwinism, and the psychology of women: A study in social myth. *American Psychologist, 30*(7), 739–754. https://doi.org/10.1037/h0076948

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stewart, L. A., & Parmar, M. K. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *The Lancet, 341*(8842), 418–422. https://doi.org/10.1016/0140-6736(93)93004-K

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions, 25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., & Stewart, L. A. (2012). Statistical analysis of individual participant data meta-analyses: A comparison of methods and recommendations for practice. *PloS ONE, 7*(10), e46042. https://doi.org/10.1371/journal.pone.0046042

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS ONE, 8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6*, 189. https://doi.org/10.3389/fpsyg.2015.00189

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

Unger, R. K. (1979). Toward a redefinition of sex and gender. *American Psychologist, 34*(11), 1085–1094. https://doi.org/10.1037/0003-066X.34.11.1085

Valentine, J. C., Biglan, A., Boruch, R .F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*, 103–117. https://doi.org/10.1007/s11121-011-0217-6

van Anders, S. M. (2015). Beyond sexual orientation: Integrating gender/sex and diverse sexualities via sexual configurations theory. *Archives of Sexual Behavior, 44*, 1177–1213. http://dx.doi.org/10.1007/s10508-015-0490-8

van Walraven, C. (2010). Individual patient meta-analysis—rewards and challenges. *Journal of Clinical Epidemiology, 63*(3), 235–237. https://doi.org/10.1016/j.jclinepi.2009.04.001

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

Wai, J., Hodges, J., & Makel, M. C. (2018). Sex differences in ability tilt in the right tail of cognitive abilities: A 35-year examination. *Intelligence, 67*, 76–83. https://doi.org/10.1016/j.intell.2018.02.003

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

Watt, H. M., Bucich, M., & Dacosta, L. (2019). Adolescents' motivational profiles in mathematics and science: Associations with achievement striving, career aspirations and psychological wellbeing. *Frontiers in Psychology, 10*, 990. https://dx.doi.org/10.3389/fpsyg.2019.00990

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research.* Sage Publications.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1*(2), 125–151. https://doi.org/10.1177/0891243287001002002

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In A. J. Elliot, *Advances in motivation science* (Vol. 7, pp. 161-198). https://doi.org/10.1016/bs.adms.2019.05.002

Wigfield, A., Eccles, J. S., & Möller, J. (2020). How dimensional comparisons help to understand linkages between expectancies, values, performance and choice. *Educational Psychology Review.* https://doi.org/10.1007/s10648-020-09524-2

Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon & R. M. Lerner (Series eds.) and N. Eisenberg (Vol. ed.), *Handbook of child psychology. Vol. 3: Social, emotional, and personality development* (6th ed., pp. 933–1002). Wiley.

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In J. M. Olson & M. P. Zanna, *Advances in experimental social psychology* (Vol. 46, pp. 55–123). Academic Press.

World Economic Forum. (2015). *New vision for education. Unlocking the potential of technology.* http://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist, 70*(1), 10–20. https://doi.org/10.1037/a0038208

2

Study I

Top-Performing Math Students in 82 Countries: A Meta-Analysis of Gender Differences in Achievement, Achievement Profiles, and Achievement Motivation

Keller, L., Preckel, F., Eccles, J. S., & Brunner, M. (2020). *Top-Performing Math Students in 82 Countries: A Meta-Analysis of Gender Differences in Achievement, Achievement Profiles, and Achievement Motivation.* Manuscript submitted for publication. The manuscript has been posted as a preprint on PsyArXiv (https://psyarxiv.com/73wap).

2 Study I:

Top-Performing Math Students in 82 Countries: A Meta-Analysis of Gender Differences in Achievement, Achievement Profiles, and Achievement Motivation

Abstract

This meta-analysis examined gender differences in achievement, achievement profiles, and achievement motivation in mathematics, reading, and science among 115,481 top-performing adolescent math students (top 5% in their respective countries). In the top 5% in mathematics, male students were overrepresented (female-to-male ratio 1:1.50). Furthermore, female students possessed better reading skills (d = -0.23) and more positive reading attitudes (-0.64 ≤ d ≤ -0.38). Male students had stronger math self-efficacy (d = 0.32) and demonstrated mathematics-oriented achievement profiles, whereas female students' profiles were more balanced across domains. Female students were more interested in organic and medical fields (-0.44 ≤ d ≤ -0.30), whereas male students showed greater interest in physics-related topics (0.39 ≤ d ≤ 0.54). Gender equality indicators moderated the proportion of female students in the top 5% in mathematics and explained variability in achievement profiles. Results are explained by social-role-theory and expectancy-value-theory; implications for women's underrepresentation in STEM are discussed.

Keywords: gender differences, achievement, achievement motivation, top-performers, meta-analysis

Top-Performing Math Students in 82 Countries: A Meta-Analysis of Gender Differences in Achievement, Achievement Profiles, and Achievement Motivation

Understanding the underrepresentation of women in math-intensive fields such as science, technology, engineering, and mathematics (STEM) remains a concern of scientists and society (e.g., Halpern et al., 2007). For example, in 2016, across all member states of the European Union (Eurostat, n.d.), the percentage of male students was 74% among all students in the fields of engineering, manufacturing, and construction-related studies. By contrast, the percentage of female students was 71% among all students in fields related to health and welfare and 78% in the field of education. Similar results were reported in the U.S. (National Science Board, 2016).

In their influential review, Ceci et al. (2014) concluded that future research should focus on the "barriers to women's full participation in mathematically intensive academic science fields [that are] rooted in pre-college factors and the subsequent likelihood of majoring in these fields" (p. 76). Several pre-college factors contribute to women's underrepresentation in STEM (e.g., Ceci et al., 2009; Wang & Degol, 2013, 2017). In particular, previous research showed that those most likely to major in and enter STEM fields are top-performing math students (Halpern et al., 2007; Lubinski & Benbow, 2006; Park et al., 2007). Thus, gender differences in the highest levels of achievement or in the right tail of the achievement distribution in math are vital for explaining gender disparities in STEM (Ceci et al., 2009, 2014; Halpern et al., 2007). Furthermore, educational and occupational choices are shaped by students' achievement profiles (e.g., Park et al., 2007; Wang et al., 2013). Individuals whose specific strength is mathematics will accomplish more professionally in STEM fields than individuals with equivalent math skills but also high verbal skills (Park et al., 2007). Finally, students' achievement motivation predicts their educational and occupational choices (e.g., Eccles, 1994; Halpern et al., 2007). Students most likely enroll in

courses and pick occupations that they think they can master and for which they experience a high task value (Eccles, 1994). To sum up, differences in (a) the proportions of top-performing male and female students in mathematics but also gender differences in (b) their achievement, (c) achievement profiles, and (d) their domain-specific achievement motivation will most likely contribute to gender differences in STEM (Ceci et al., 2009; Halpern et al., 2007; Wang & Degol, 2013, 2017). However, no meta-analysis has ever investigated these pre-college factors simultaneously in samples of top-performing math students. The overarching goal of the present study was therefore to provide a uniquely comprehensive meta-analysis that has been missing from the field of mathematical talent, gender, and STEM. First, previous research on gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation has focused on U.S. samples and used data from special programs with unknown representativeness for the full student population (e.g., Lubinski & Benbow, 2006). In our search for available databases appropriate for addressing our research objectives, we therefore focused on representative, unselective samples from well-defined populations, an approach considered the "gold standard" (Hedges & Nowell, 1995; Reilly et al., 2019). Second, capitalizing on these data, we examined gender differences in achievement, achievement profiles, and achievement motivation[3] in top-performing math students from 82 countries in three core academic domains: mathematics, reading, and science. Previous meta-analyses that investigated gender differences in top-performing math students focused on a single domain (mostly mathematics; e.g., Hyde, Fennema, & Lamon, 1990; Lindberg et al., 2010), which precluded an examination of gender differences in achievement profiles across domains. Third, we investigated several possible moderator variables that may help explain why gender differences in top-performing math students are

[3] Unless otherwise indicated, we use the term "gender differences in top-performing math students" to indicate "gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation" for the remainder of the article.

larger in some countries than in others. In particular, drawing on social role theory (Wood & Eagly, 2012) and expectancy-value theory (Eccles et al., 1983), we examined the extent to which gender equality indicators are related to the cross-national heterogeneity of gender differences in top-performing math students. In sum, we provide an authoritative synthesis of gender differences in multiple key (pre-college) factors in top-performing math students that previous research has identified to explain gender differences in STEM.

The Development of Gender Differences

Various theoretical explanations for the emergence of gender differences have been offered while simultaneously acknowledging that biological, psychological, and environmental factors constantly interact in reciprocal feedback loops to shape individuals' development (Miller & Halpern, 2014). In the following, we focus on the role of sociocultural factors for the development of gender differences.

The Role of Sociocultural Factors in the Development of Gender Differences

Social role theory (Eagly, 1987; Wood & Eagly, 2012) and expectancy-value theory (Eccles et al., 1983) provide prominent psychological models for explaining why sociocultural factors influence gender differences in the development of mathematical talent. Both theories propose that gender differences emerge because of differences in males' and females' roles in society. These differences in gender roles are based on gender stereotypes (Eccles, 1994; Wood & Eagly, 2012), that is, on beliefs about differences between females and males (Ashmore & Del Boca, 1979). Social role theory explains the psychological mechanisms that lead to gender stereotypes and how gender-typed roles influence gender differences in achievement and motivation, whereas expectancy-value theory is more focused on explaining the latter aspect.

Social role theory (Eagly, 1987; Wood & Eagly, 2012) argues that gender stereotypes emerge because people infer that there is a correspondence between people's external

behavior and their internal characteristics (Wood & Eagly, 2012). For example, because women tend to do domestic work and communally demanding jobs, people infer that women are warm, caring, and socially skilled. Similarly, because men tend to take on strength-intensive roles and high-status roles, people infer that men are assertive, dominant, and forceful (Wood & Eagly, 2012). In addition to this descriptive aspect of gender roles, gender roles also have a prescriptive function. By considering these typical attributes as generally desirable and admirable for each sex, gender role beliefs promote norms and socialization practices (e.g., by parents, teachers, and peers) that encourage children to acquire the skills, characteristics, and preferences that support their society's division of labor. Over time, gender roles tend to be internalized as gender identities and thus facilitate stereotype-consistent behavior through self-regulatory processes (Eagly, 1987; Wood & Eagly, 2012). Empirical research supports that there is a direct link between gender stereotypes and the representation of men and women in social roles (Koenig & Eagly, 2014). In educational contexts, research indicated that a higher female enrollment in tertiary science education and the representation of women in the science workforce were related to weaker national gender–science stereotypes (Miller et al., 2015). Further, if the gender roles that women are expected to fulfill within a society do not include math- or science-related activities, female students may encounter social barriers (e.g., mathematics and science are stereotyped as male domains) and perhaps even structural barriers (e.g., girls are disadvantaged in terms of formal access to [math or science] education). In turn, this can impair girls' development of skills in mathematics or science. For example, a large-scale study by Nosek et al. (2009) demonstrated that gender differences in mathematics and science were larger in countries where residents hold stronger stereotypes that associate men with science.

Expectancy-value theory (Eccles et al., 1983) proposes more specifically that gender differences in achievement and motivational aspects, such as domain-specific expectations for

success (e.g., self-efficacy beliefs, self-concepts) and values (e.g., interest, enjoyment, emotional costs such as anxiety), emerge through the processes by which children are socialized into gender roles. For example, the meta-analysis by Lytton and Romney (1991) found differential parental encouragement of gender-typed activities. Hence, parents' socialization processes may differ for boys and girls with the consequence that parents provide different learning opportunities and experiences to boys and girls. For instance, mothers seem to provide more math-supportive environments for boys than for girls by buying more math-related toys for their sons than for their daughters (Jacobs et al., 2005). Ultimately, expectancy-value theory states that such socialization processes will result in gender differences in male and female students' domain-specific achievement and motivation, which in turn lead to gender differences in educational and occupational preferences and choices.

To conclude, both expectancy-value theory and social role theory emphasize that gender-typed socialization processes are sociocultural factors that influence the development of gender differences. According to both theories, gender differences should be smaller in societies that endorse gender-typed roles to a smaller extent and that have greater gender equality. More specifically, expectancy-value theory predicts that if a female student gender-types a domain such as mathematics as masculine and not in line with her gender role values, she is less likely to value mathematics and less likely to put effort into math-related fields, especially if she does not expect to do well. Consequently, she is more likely to perform poorly in mathematics and to avoid choosing math-related studies and careers (Eccles, 1994; Meece et al., 1982). Thus, socialization processes that are more gender-typed produce larger gender differences in achievement and motivation.

Several studies that have investigated the link between sociocultural factors and gender differences focused on gender equality as an important sociocultural factor. Prior

cross-national studies have primarily analyzed the role of gender equality for gender differences at the level of the general student population (Baker & Jones, 1993; Else-Quest et al., 2010; Guiso et al., 2008; Reilly, 2012; Riegle-Crumb, 2005; Stoet & Geary, 2013, 2015). Only a few studies have investigated the role of gender equality for gender differences at the right tail of the ability distribution. Using data from TIMSS 1995 (Penner, 2008) and PISA 2003 (Guiso et al., 2008), two studies found that the proportion of female students in the top 5% in mathematics increased as gender equality in a country increased. Similar to results on the population level, findings varied to some extent depending on the gender equality indicators used (Penner, 2008). Furthermore, Hyde and Mertz (2009) found that the percentage of female students on a country's International Mathematical Olympiad team was significantly correlated with its Global Gender Gap Index (GGI). Thus, there is evidence that sociocultural factors may affect the development of mathematical talent: In countries with higher levels of gender equality, more female students score at the highest levels of math achievement. Yet, it is unknown whether sociocultural factors are also related to gender differences (e.g., in math, reading, or science achievement) *within* the group of top-performing math students.

Gender Differences in Achievement and Achievement Motivation

The evidence on gender differences in achievement and achievement motivation can be divided into mean differences (at the midpoint of a distribution for the general student population) and differences in the right tail (e.g., the top 10%, 5%, and 1%; Ceci et al., 2009, 2014; Halpern et al., 2007). The latter is typically considered to be more relevant for improving the understanding of gender disparities in STEM because students from the right tail are most likely to major in and enter STEM fields (Ceci et al., 2009, 2014; Halpern et al., 2007). Nevertheless, results obtained for the general population are still valuable as they often

predict trends in the right tail (Ceci et al., 2009) and they offer a way to benchmark results as obtained for top-performing individuals.

As evident from Figure 1 (see also Tables S1 in the Supplemental Online Material [SOM] of Study I), there are numerous meta-analyses and large-scale studies which provide strong empirical evidence on gender differences in students' achievement in mathematics, reading, and science in the general population. Considerably fewer meta-analyses and large-scale studies have examined gender differences in students' achievement motivation in the general population (Figures 1 and 2; Table S2). Most importantly, Figures 1 and 2 show that the evidence base is particularly weak for gender differences in top-performing math students. Only three meta-analyses examined gender gaps in math performance in this group of students (Baye & Monseur, 2016; Hyde, Fennema, & Lamon, 1990; Lindberg et al., 2010). The results indicated that gender differences in favor of male students are somewhat larger in highly selected samples (e.g., the top 5%) than in the general population ($0.15 \leq d \leq 0.54$ as compared to $-0.05 \leq d \leq 0.31$ for the general population; see Table S1). However, some of these findings are at least partially based on data from talent search studies[4] (Hyde, Fennema, & Lamon, 1990; Lindberg et al., 2010). A study that used representative data exclusively from unselected top-performing math students reported a smaller gender gap in math achievement ($d = 0.15$; Baye & Monseur, 2016). Notably, there is evidence that gender differences in math achievement in top-performing math students vary cross-nationally (Stoet & Geary, 2013).

In addition, studies have revealed a substantial overrepresentation of male students among top-performers in mathematics. Two meta-analyses that analyzed representative data sets from the US reported a female-to-male ratio of 1:1.50 to 1:4.09 in the top 5% in

[4] Participants in talent search studies represent a selected student group, particularly in that they are aware of their ability because of their selection into the program. This awareness most likely influences their self-beliefs, motivation, and possibly also their performance.

mathematics (Hedges & Nowell, 1995; Nowell & Hedges, 1998). Studies that used more recent data sets from representative international large-scale assessments (Machin & Pekkarinen, 2008; Stoet & Geary, 2013) and state or national assessments from the US (Hyde et al., 2008; Reilly et al., 2015) also found a preponderance of male students in the top 5% in mathematics. However, the female-to-male ratios were comparatively low and varied across countries (1:1.09 to 1: 2.13). Research findings from talent search programs showed that the female-to-male ratio in the top 0.5% of math ability rapidly declined from the early 1980s (1:2.61) to the early 2010s (1:1.37; SMPY; Makel et al., 2016). Within the top 0.01% of math ability in the SMPY, the decline was even sharper (Makel et al., 2016; Table 2). Using a different U.S. talent search database, Olszewski-Kubilius and Lee (2011) reported slightly higher female-to-male ratios (1:2.5 to 1:3.7) for students in the top 2% in mathematics between 2000 and 2008 (compared with the results from Makel et al., 2016).

Regarding top-performing math students' achievement motivation, there is only one meta-analysis that covered gender differences in math motivation. The findings suggested that gender differences in math anxiety were negligible in highly selected samples (Hyde, Fennema, Ryan et al., 1990). As the overview in Figures 1 and 2 illustrates, for top-performing math students, there are neither meta-analyses or large-scale studies that have examined gender differences in reading or science achievement nor in the achievement motivation in science or verbal domains.

Gender Differences in Achievement Profiles

Achievement profiles are comprised of the pattern and structure of achievement in multiple domains within an individual. One way to create achievement profiles is to calculate achievement tilts by subtracting a student's test score in one domain from the same student's test score in another domain (e.g., Wai et al., 2018). Previous research has demonstrated that achievement tilts in math and verbal domains at the age of 16 (Dekhtyar et al., 2018) and on

college entrance exams (Coyle et al., 2014, 2015; Wang et al., 2013, 2017) predicted career choices in adulthood in the general population. Math tilts were associated with STEM majors (e.g., science and math) and STEM careers, whereas verbal tilts were associated with humanities majors (e.g., English and history) and humanities careers (Coyle et al., 2014, 2015; Dekhtyar et al., 2018; Wang et al., 2013, 2017).

Female and male students in the general population have been found to differ in their achievement profiles such that male students were more likely to show math tilts (and STEM preferences), whereas female students were more likely to show verbal tilts (and humanities preferences; Coyle et al., 2014, 2015; Dekhtyar et al., 2018; Wang et al., 2013). Ability tilts seem to be larger in the right tail of the ability distribution than in the general population (Lohman et al., 2008). Similar to findings in the general population, achievement tilts in high-ability students predicted their educational and career choices. Students in talent search samples who scored higher on math relative to verbal achievement at the age of 13 gravitated toward STEM occupations; however, students who scored higher on verbal relative to math achievement gravitated toward the humanities (Lubinski et al., 2001; Park et al., 2007). In addition, there is evidence of gender differences in achievement tilts in high-achieving students. Wai et al. (2018) examined gender differences in math and verbal achievement tilts in academically talented students in the US across 35 years and found that more male than female students showed positive math tilts and more female than male students showed positive verbal tilts. Furthermore, gender differences in achievement tilts increased with achievement level (i.e., from the top 5% to the top 1% to the top 0.01% of ability; Wai et al., 2018). However, there are no meta-analyses or large-scale studies that have investigated gender differences in achievement profiles in the general population or in top-performing math students (see Figures 1 and 2).

The Present Study

The present meta-analysis had two main research goals. The first goal was to provide reliable and widely generalizable empirical knowledge about gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in three core academic domains: mathematics, reading, and science. Importantly, our meta-analysis focused on students in secondary school because educational and occupational choices leading to STEM careers are shaped by pre-college factors during adolescence (Ceci et al., 2009, 2014; McDaniel, 2016). To this end, we capitalized on international, representative, and unselected individual participant data from well-defined populations of students at the end of compulsory education. In doing so, the present study is the first to meta-analyze important gender differences in top-performing math students' reading and science achievement, achievement profiles (i.e., math–reading, science–reading, and math–science profiles), and achievement motivation related to mathematics, reading, and science. Furthermore, we significantly extended the findings on the proportion of female students in the group of top-performing math students and on gender differences in math achievement in this group of students (e.g., Guiso et al., 2008; Penner, 2008) by using more recent data from a substantially larger number of countries.

The second goal of this meta-analysis was to investigate the moderating roles of gender equality for gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation. Specifically, we used domain-specific gender equality indicators (e.g., the tertiary enrollment ratio or women's share of research positions) to examine the specific processes that might lead to the observed gender gaps (Else-Quest & Grabe, 2012). According to social role theory and expectancy-value theory, we expected that gender differences in top-performing math students would decrease with increasing levels of gender equality. Furthermore, we predicted that the share of female

students in the top 5% in mathematics would increase with increasing levels of gender equality.

Method

Identification of International Large-Scale Assessments

To meta-analyze gender differences in top-performing math students, we sought for individual participant data from international large-scale assessments rather than summarizing published results obtained from empirical studies (e.g., different national large-scale assessments). Our reasons for doing so were twofold. First, a key feature of international large-scale assessments is the applied standardization protocol for all phases of the study (e.g., applying the same set of standardized measures in all participating countries). Thus, using individual participant data from international large-scale assessment data allowed us to control for several sources of unwanted heterogeneity in effect sizes (for details see S1 in the SOM), which, in turn, substantively improved the reliability, precision, and statistical power of the meta-analytic syntheses (Valentine et al., 2010) and moderator analyses (Hempel et al., 2013). Second, international large-scale assessments examine representative, unselected student samples in many different countries. Thus, these data naturally support the wide generalization of findings on gender differences in top-performing students within and across countries.

To identify all potential international large-scale assessments we conducted a systematic search, which is described in detail in section S1 in the SOM. Figure 3 provides an overview of the selection process and the number of publications and unique international large-scale assessments identified in each step. After evaluation, only the Programme for International Student Assessment (PISA) met all the inclusion criteria. Thus, for the present meta-analysis, all available individual student data from six PISA cycles were used (i.e.,

samples between 2000 and 2015). Protocols can be accessed via the Open Science Framework (OSF; Soderberg, 2018) at https://osf.io/jnxwm/.

Description of the PISA Study and Study Participants

PISA is a triennial international survey conducted by the Organisation for Economic Co-operation and Development (OECD). It is aimed at evaluating education systems worldwide at the end of compulsory education by testing the skills and knowledge of 15-year-old students in the key domains of mathematics, reading, and science. PISA capitalizes on a two-stage stratified sampling design to achieve representative probability samples (a detailed description of the sampling procedures can be found in section S2 in the SOM). Consequently, PISA results can be generalized to the larger population of 15-year-old students in all participating countries.[5] Across all PISA cycles, a total of 2,280,502 students from 83 countries participated. In the present study, students who belonged to the top 5% in mathematics in their country in a certain PISA cycle were defined as top-performers in mathematics. Table S3 in the SOM presents the country-specific cut-off values for the top 5% in mathematics for each PISA cycle.

Students from Liechtenstein were excluded from the present analyses due to the small number of students who scored among the top 5% in mathematics in all PISA cycles in Liechtenstein ($n < 30$). In the year 2000, nine students were excluded from the analyses because they were missing information on their gender. Of note, in PISA 2000, a mathematics achievement score was provided in the public use file for only 56% of the students (for an even smaller proportion of students, science achievement scores were available). This resulted in a smaller number of mathematically top-performing students compared with the other PISA cycles where a mathematics achievement score was provided for all students. The final

[5] Students from OECD and non-OECD countries or economic regions can participate in PISA. For ease of presentation, we refer to both "countries" and "economic regions" as "countries" in this article.

sample included data from 82 countries (Figure S1 in the SOM) and 115,481 top-performing math students (see Table 2 for the sample characteristics).

Measures

Achievement

In PISA cycles 2003 to 2012, mathematical achievement was assessed in four categories: change and relationships, space and shape, quantity, and uncertainty and data. In PISA 2000, the assessment covered just two categories: (a) space and shape and (b) change and relationships. Reading literacy was assessed in three different categories: the abilities to access and retrieve information, integrate information and interpret texts, and reflect upon and evaluate texts. In 2006, reading literacy results were not reported for the US because of an error in the printing of the test booklets. Science literacy was assessed in the categories physical systems, living systems, earth and space systems, technology systems, scientific inquiry, and scientific explanations. See section S3 in the SOM for further details on the achievement measures.

Achievement Motivation

To assess students' achievement motivation, students reported their motivation with respect to mathematics, reading, and science on 26 self-report scales measuring, for example, their self-concept, self-efficacy beliefs, and instrumental and future-directed motivation, anxiety, interest, and enjoyment. Students gave their answers in a forced-choice format (for math intentions) and on 4-point rating scales (for the remaining 25 measures). Tables S4 to S7 in the SOM provide an overview of the scales, the corresponding items, the response options for the items, and the scale score reliabilities (i.e., internal consistencies).

Gender Equality Indicators as Moderators

We selected widely regarded specific measures of gender equality recommended by Else-Quest and Grabe (2012) in the areas of education (i.e., gender ratios in primary, secondary,

and tertiary education enrollment) and higher positions (i.e., women's share of higher positions and research positions in a country) that are theoretically relevant as moderators of girls' and women's engagement in mathematics. Gender equality in education reflects the valuation of female students' education in a society, and gender equality in higher positions reflects the penetration of the glass ceiling (Else-Quest & Grabe, 2012). Table 2 lists and provides descriptions of the indicators used in the present study. In the moderator analyses, we only included data on the specific gender equality indicators that were assessed in the same assessment years as the PISA cycles (e.g., the primary enrollment ratio has been reported annually since 1970, but we only included the data from 2000, 2003, 2006, 2009, 2012, and 2015 in the present study). To maximize the number of countries with data on women's share of research positions, we combined data from the OECD and the United Nations Educational, Scientific, and Cultural Organization (UNESCO). When data were available from only one data set, we used that information. When data were available from two sources, we averaged them. Values for all moderators, their intercorrelations, and descriptive statistics are presented in Tables S8 and S9 in the SOM. The intercorrelations among moderators showed that all moderators contained unique information.

Data Analysis

General Procedure

The present study is a meta-analysis of individual student data as provided in large-scale data sets (Cooper & Patall, 2009; see Hedges & Nowell, 1995, for an application) that we conducted in three steps in accordance with the analysis strategy proposed by Cheung and Jak (2016) for big data. In the first step, we computed effect sizes using the individual student data for each country and each PISA cycle. In the second step, we meta-analyzed the effect sizes to estimate (a) the average effect sizes for gender differences in achievement, achievement profiles, and achievement motivation and (b) the heterogeneity of effect sizes

within and between countries. In the third step, we examined the extent to which moderator variables may explain the observed heterogeneity in effect sizes. Analyses were conducted using the statistical software R (version 3.6.1; R Core Team, 2019). The R code for reproducing the results and figures from the present study can be found on the OSF.

Step 1: Effect Size Computation. We analyzed the country-specific magnitude of gender differences in achievement and achievement motivation by computing the effect size *d* (Cohen, 1988). Cohen's *d* is the effect size for the standardized mean difference between two groups on a continuous variable (e.g., the mean difference between male and female students on a continuous measure of mathematics achievement). Thus, country-specific *d* was computed, with $d = (M_m - M_f)/SD_{OECD}$, M_m = the mean for male students, M_f = the mean for female students, and SD_{OECD} = the standard deviation of the total student sample from the OECD countries. Hence, positive values indicated an advantage of male students and negative values an advantage of female students. In accordance with Hyde (2005), we defined five ranges of effect sizes: negligible ($0.00 < |d| \leq 0.10$), small ($0.10 < |d| \leq 0.35$), moderate ($0.35 < |d| \leq 0.65$), large ($0.65 < |d| \leq 1.00$), and very large ($|d| > 1.00$).

To examine achievement profiles, we subtracted an individual student's achievement score in one domain from this student's achievement score in another domain, resulting in three different profiles: math–reading, science–reading, and math–science. Overall, we computed three effect sizes to capture gender differences in achievement profiles: country-specific mean profile scores for male and female students, the gender-specific percentage of tilts within each profile score, and the percentage of nonoverlap in gender-specific profile distributions. Nonoverlaps of 8%/24%/41%/55% can be considered to represent small/medium/large/very large effects, respectively.

Further information on the calculation of effect sizes can be found in section S4, country-specific effect sizes for each outcome can be found in Tables S10 to S20 in the SOM, and standard errors for all unweighted effect sizes can be accessed via the OSF.

Step 2: Meta-Analysis. To meta-analyze the effect sizes, we used the R package "metaSEM" (version 1.2.2; Cheung, 2015) that implements random-effects models with maximum likelihood estimation to allow the true effect to vary (Borenstein et al., 2009; Cheung, 2015). When effect sizes were available only for a single PISA cycle, we used two-level random effects models. In the two-level random effects models, variance estimates for the various effect sizes (as obtained in Step 1) defined Level 1; Level 2 captured variability in effect sizes between countries. When effect sizes were available for several PISA cycles, we used three-level random effects models to account for the dependencies between the effect sizes (i.e., effect sizes obtained for several PISA cycles within countries). In the three-level random effects models, variance estimates for the various effect sizes (as obtained in Step 1) defined Level 1. Level 2 captured variability in effect sizes between PISA cycles within countries, and Level 3 captured variability in effect sizes between countries. We computed three statistics to assess the heterogeneity of effect sizes: T, I^2, and Q (Borenstein et al., 2009). T is the standard deviation of the effect size parameters (Borenstein et al., 2009). I^2 represents the proportion of observed heterogeneity that is real and not due to random noise and has a range of 0% to 100% (Higgins & Thompson, 2002). For the three-level models, we estimated T and I^2 within countries (T_{L2}, I^2_{L2}), between countries (T_{L3}, I^2_{L3}), and in total ($T_{\text{total}} = \sqrt{T^2_{L2} + T^2_{L3}}$, $I^2_{total} = I^2_{L2} + I^2_{L3}$); for the two-level models, we estimated T_{total} and I^2_{total} only. The Q test statistic (introduced by Cochran, 1954) is computed by summing the squared deviations of each individual effect size estimate from the corresponding average effect estimate where individual effect sizes are weighted by their sampling variance (Huedo-Medina et al., 2006). A statistically significant value of Q is typically taken to indicate effect

size heterogeneity. We considered all three statistics to evaluate the variability of effect sizes and, consequently, to decide whether it would be appropriate to conduct further moderator analyses. Specifically, moderator analyses were performed if the Q statistic associated with a certain effect size was significant (Lipsey & Wilson, 2001) or if T_{total} or the I^2_{total} indicated at least moderate heterogeneity. Whereas there are established guideline values for moderate I^2 values ($I^2 \geq 30\%$, Higgins & Green, 2011), these guideline values are lacking for T. Hence, to assess which T_{total} value can be considered moderate, we computed empirical benchmark values using data on standardized mean differences (i.e., Cohen's d and Hedges' g) provided by van Erp et al. (2017). Cut-off scores were based on the approach presented by Hemphill (2003) and Bosco et al. (2014). Thus, T_{total} values in the middle third ($.12 \leq T_{total} < .28$) could be considered to indicate a moderate level of heterogeneity.

Step 3: Mixed-Effects Models and Moderator Analysis. The mixed-effects meta-analysis extends the random-effects meta-analysis by explaining the heterogeneity of the effect sizes within and between countries by moderator variables (Borenstein et al., 2009; Cheung, 2015). We ran multivariate meta-regression models for each effect size (i.e., the dependent variable) using the following set of moderator variables (i.e., the independent variables): women's share of higher positions, women's share of research positions, and enrollment ratios in primary, secondary, and tertiary education.

As in almost every meta-analysis, some data on moderator variables were missing (i.e., gender equality indicators were not available for all countries). We therefore followed the recommendations by Pigott (2019) and Tipton et al. (2019) and used multilevel multiple imputation (e.g., Grund et al., 2018) to estimate unreported values and to account for the clustered data structure (effect sizes nested in countries). To facilitate the interpretation of the results, moderator variables that represented ratios were log-transformed before multiple imputation and then used in the meta-regression. Subsequently, the regression coefficients

were divided by 100 such that a 1% increase in the moderator variable increased (or decreased) the dependent variable by coefficient/100 units. Further details on the data analysis are provided in section S4 in the SOM.

Analysis for Possible Bias

Meta-analyses using individual participant data (e.g., PISA public use files) are generally considered the most reliable approach for synthesizing data (e.g., Stewart & Tierney, 2002). Nevertheless, there remains a potential concern that data and results are affected by various sorts of bias (Ahmed et al., 2012) that—if present—could imply that the magnitude and heterogeneity of gender differences in top-performing math students may be under- or overestimated. We provide a detailed description of our analysis for possible bias in the section S5 in the SOM. We conclude that most sources of bias (reviewer selection bias, publication-related bias) are minimized in the present study. However, the reach of this study is limited because of restrictions of the country sample (not all countries worldwide participate in PISA; see also the Discussion section and section S5 in the SOM).

Results

Proportions of Male and Female Students in the Top 5% in Mathematics

The overall percentage of female students in the top 5% in mathematics, averaged across all studies, was 40% (Table 3), corresponding to a female-to-male student ratio of 1:1.50. Figure 4 shows the distribution of the percentages of female students. Given the heterogeneity in the effect sizes (Table 3), we conducted analyses for moderator variables to explain the heterogeneity in effect sizes. Tertiary enrollment ratios positively predicted the proportion of female students in the top 5% in mathematics (b = 0.04), indicating that a rise in the tertiary enrollment ratio by 1% was associated with an increase in the percentage of female students in the top 5% in mathematics by 0.04% (under control the other gender equality indicators; Table 4). That is, the larger the percentage of female students enrolled in a university

compared with the percentage of male students enrolled, the larger the proportion of female students in the group of top-performing math students.

Gender Differences in Achievement

The overall weighted mean effect size of the gender difference in mathematics achievement was $d = 0.05$ (Table 3), representing a negligible gender difference in top-performing math students. Figure 5A shows that the range of effect sizes was narrow and that gender differences were negligible in almost all countries. Because effect sizes were homogenous (Table 3), moderator analyses were not performed.

The overall weighted mean effect size of the gender difference in reading achievement was $d = -0.23$ (Table 3), indicating that, on average, female students had better reading performance than male students did. Figure 5A shows that the magnitude of effect sizes varied across studies with the vast majority of effect sizes indicating that female students outperformed their male counterparts in reading. The heterogeneity measures showed that gender differences in reading achievement were heterogeneous (Table 3). We therefore conducted further moderator analyses, but the gender equality indicators did not significantly explain the variation in effect sizes (Table 4).

The overall weighted mean effect size of the gender difference in science achievement was $d = 0.01$ (Table 3), showing that male and female students performed similarly in science. Figure 5A displays the distribution of gender differences in science and displays that almost all effect sizes were negligible or small. Because the heterogeneity measures indicated that the effect sizes were heterogeneous, moderator analyses were conducted. Yet, none of the moderator variables significantly predicted variability in effect sizes (Table 4).

Gender Differences in Achievement Profiles

In the math–reading profile, mathematically top-performing male students' math achievement clearly exceeded their reading achievement by, on average, 57.65 points (Table 5). Although

female students' math achievement also exceeded their reading achievement, on average, by 22.71 points, the difference between mathematics and reading achievement was less pronounced for female students than it was for male students. This pattern is also displayed in Figure 6A, showing that male students gravitated toward a strongly mathematics-oriented profile, whereas female students' achievement profiles were somewhat more evenly distributed across the math–reading dimension. Female and male students' math–reading profile distributions had, on average, a nonoverlap of 44%, representing a large effect (Table 5; see Figure 6B for the distribution of effect sizes). Of all male students, 87% scored higher in mathematics than in reading (i.e., 87% demonstrated a math $tilt_{M–R}$), whereas 66% of all female students showed stronger achievement tilts in mathematics than in reading (i.e., 66% demonstrated a math $tilt_{M–R}$; Table 5).

In the science–reading profile, mathematically top-performing male students performed better in science than in reading, showing a profile score difference of, on average, 32.20 points in favor of science (Table 5), whereas female students performed almost as well in reading as in science, demonstrating a profile score difference of, on average, 2.08 points in favor of science. Figure 6A shows that male students gravitated toward a strongly science-oriented profile, whereas female students' achievement profiles were more evenly distributed over the science–reading dimension. Female and male students' science–reading profile distributions showed a mean nonoverlap of 42%, representing a large effect (Table 5; see Figure 6B for the distribution of effect sizes). Of all mathematically top-performing male students, 76% showed better achievement in science than in reading (i.e., 76% demonstrated a science $tilt_{S–R}$), and 48% of all mathematically top-performing female students demonstrated higher achievement in science than in reading (i.e., 48% demonstrated a science $tilt_{S–R}$; Table 5).

In the math–science profile, both male and female students performed better in mathematics than in science, demonstrating profile score differences of, on average, 24.55 and 19.87 points, respectively (Table 5). Figure 6A shows that male and female students' achievement profiles were somewhat tilted toward mathematics. Female and male students' math–science profile distributions had a nonoverlap of 18% (Table 5; see Figure 6B for the distribution of effect sizes), and thus, gender differences in the math–science profile were small. For 69% of all male students, mathematics was their strongest skill compared with science (i.e., 69% demonstrated a math $tilt_{M–S}$). Similarly, 65% of all female students scored higher in mathematics than in science (i.e., they demonstrated a math $tilt_{M–S}$; Table 5).

Moderator analyses were performed to investigate whether the heterogeneity in female and male students' profile scores, the percentage of nonoverlap between their distributions of profile scores, and the percentage of female and male students demonstrating a certain tilt in their profile scores (Table 5) could be predicted by gender equality indicators. The results in Table 6 show that enrollment ratios in tertiary education were associated with female and male students' math–reading profile scores (b_{female} = -0.15, b_{male} = -0.12) and female students' science–reading profile scores (b_{female} = -0.10). The findings indicate that when the percentage of female students enrolled in tertiary education compared with the percentage of male students enrolled in tertiary education is higher, (a) the difference between math and reading scores is smaller for female and male math top-performers and (b) the difference between science and reading scores is smaller for female math top-performers. Furthermore, women's share of research positions predicted the variation in female students' math–science profile scores. That is, the higher women's share of research positions, the smaller the difference between students' math and science scores (b_{female} = -0.98).

Gender Differences in Achievement Motivation

With regard to top-performing math students' math motivation, male students reported higher math self-efficacy than female students did. That is, compared with female students, male students reported feeling more confident about solving a specific math task (d = 0.32; Table 3). Similarly, male students reported higher intentions to focus on math than female students did (i.e., male students reported a higher intention to choose additional math courses in school and beyond compared with additional language or science courses; d = 0.27). Furthermore, male students reported, on average, higher instrumental math motivation (d = 0.16), higher math self-concept (d = 0.15), and greater interest in math (d = 0.10) than female students did. Female students reported higher self-responsibility for failure in mathematics (d = -0.13), higher math anxiety (d = -0.15), and a higher math work ethic (e.g., preparing thoroughly, paying attention in class, positive learning behavior; d = -0.16) than male students did. However, the magnitude of all gender differences in math motivation was small to negligible. The distribution of effect sizes is depicted in Figure 5B. A heterogeneity analysis revealed that for six out of nine math motivation domains effect sizes were significantly heterogeneous (Table 3). However, gender equality indicators did not account for the variation in effect sizes (Table 4).

Regarding the verbal motivation of top-performing math students, female students' reports of their reading enjoyment (d = -0.64), their interest in reading (d = -0.50), and their verbal self-concept (d = -0.38; Table 3) were higher than their male counterparts' reports (all moderate effects). Importantly, across all (or almost all) countries, female students reported a higher verbal motivation (Figure 5C). Because effect sizes were heterogeneous for the enjoyment of reading (Table 3), we performed a moderator analysis for this outcome. As shown in Table 4, gender equality indicators did not account for significant variation in effect sizes.

Table 3 also shows the weighted average effect sizes of seven components of top-performing math students' science achievement motivation: self-concept, general value, enjoyment, self-efficacy, future-oriented motivation, personal value, and instrumental motivation. The distribution of effect sizes is depicted in Figure 5D. Overall, in five out of seven science achievement motivation domains, gender differences were on average negligible, indicating that among top-performing math students, male and female students' science motivation is more similar than different. Male students reported a higher science self-concept (d = 0.19, small effect) and a higher general value of science (d = 0.12, small effect) than female students did.

In contrast to top-performing math students' general motivation in science, we found gender differences in students' interest in specific science topics. Specifically, female students were more interested in human biology (d = -0.44, moderate effect; Table 3) and in learning more about diseases (d = -0.30, small effect) and plant biology (d = -0.30, small effect) than their male counterparts were. Male students were more interested in the topics motion and forces (d = 0.54), physics (d = 0.40), and energy transformation (d = 0.39); these gender differences were all moderate in size. Figure 5E depicts the distribution of effect sizes. A heterogeneity analysis revealed that the effect sizes for science self-concept, general value of science, science self-efficacy, future-oriented science motivation, enjoyment of science, instrumental science motivation, and interest in physics were heterogeneous (Table 3). However, the magnitudes of the gender differences did not depend on gender equality indicators (Table 4).

Discussion

For the group of top-performing math students, there is only little reliable and widely generalizable knowledge on gender differences in pre-college factors related to STEM, including adolescents' achievement and achievement motivation. To address this research

gap, we meta-analyzed representative individual participant data of 15-year-olds in 82 countries (PISA 2000–2015). The first goal of the present meta-analysis was to examine gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in three core academic domains—mathematics, reading, and science. The second goal was to investigate the moderating role of gender equality indicators for gender differences in this group of students.

Gender Differences in Achievement, Achievement Profiles, and Achievement Motivation

Female and male students in the top 5% in mathematics were similar in their achievement in mathematics, reading, and science and in 23 out of 30 motivational characteristics. However, the present meta-analysis provided strong empirical evidence that there are important gender differences in top-performing math students' achievement profiles, verbal motivation, and specific science interests as well as in the proportion of female students in the top 5% in mathematics. In the following, we will discuss the observed gender differences.

Gender Differences in Achievement Profiles

The present study is the first to systematically examine gender differences in achievement profiles in top-performing math students. We found that among students in the top 5% in mathematics, male students showed more distinct achievement profiles than female students did: Male students' strongest skill was more often mathematics or science than reading, whereas female students' achievement profiles were more evenly distributed across all achievement domains, especially in the math–reading and science–reading profiles. Importantly, female and male students' math–reading and science–reading profiles did not overlap much. This large lack of overlap underscores the distinctiveness of male and female students' achievement profiles in the group of top-performing math students. Similar gender differences have been found for math–verbal achievement profiles in the general population (Coyle et al., 2014, 2015; Dekhtyar et al., 2018).

Gender Differences in Verbal Motivation and Science Interests

Regarding gender differences in students' verbal motivation, we found that across (almost) all countries, mathematically top-performing female students reported higher reading enjoyment, interest, and verbal self-concept than male students. These gender gaps are similar to those found in the general population (Brunner et al., 2009; OECD, 2003, 2010; Wilgenbusch & Merell, 1999; see Figure 1 and Table S2).

Moreover, regarding top-performing math students' interest in specific science topics, female students reported greater interest in human biology, diseases, and plant biology, whereas male students were more interested in physics- and engineering-related topics, such as motion and forces, physics, and energy transformations. These results are consistent with gender differences in specific science domains found in the general population, such as in physics (d = 0.56), mechanics and electronics (d = 1.21), engineering (d = 0.83), and medical services (d = -0.40; Su & Rounds, 2015; Figure 2 and Table S2). Our results suggest that gender differences on the "things–people" dimension (i.e., that men prefer working with things or inorganic topics, and women prefer working with people or organic topics, d = 0.93, Su et al., 2009; see also Morris, 2016 and Su & Rounds, 2015) also apply for students performing at the highest levels of mathematics. Notably, this is the first time that gender differences in science motivation have been examined meta-analytically for the group of top-performing math students. Furthermore, effect sizes were mostly homogenous across PISA cycles and countries. This suggests that gender differences in interest in specific science topics are a rather universal phenomenon.

One possible explanation for female students' tilt toward biological and health interests is provided by the goal congruity model (Diekman et al., 2011). According to this model, women tend to select and pursue communal goals (i.e., working with or helping others), a tendency that is rooted in broad cultural expectations (i.e., gender roles) that are

internalized due to the rewards and punishments associated with role congruity and incongruity (Diekman et al., 2011). For example, a recent meta-analysis by Eagly et al. (2020) that investigated nationally representative U.S. public opinion polls between 1946 and 2018 showed that, on average, 85% of the respondents ascribed communal traits more to women than men and that this attribution has risen over time. Accordingly, when female students in the group of top-performing math students pursue their stronger interests in biological and health sciences, this might suggest that they perceive a better match between these areas and communal goals than they perceive for areas such as the physical sciences and engineering-related sciences (Diekman et al., 2010).

The Proportion of Female Students in the Group of Top-Performing Math Students

The present study significantly expanded the evidence base on the proportion of female students in the group of top-performing math students by meta-analyzing this proportion with a more comprehensive set of data and a considerably larger number of countries compared with previous research (i.e., Guiso et al., 2008; Penner, 2008). We found that, on average, female students were underrepresented in the top 5% in mathematics: Two out of five students (female-to-male ratio of 1:1.50) in the group of top-performing math students were female. Thus, our results fall within the range of previous studies that reported female-to-male ratios between 1:1.09 and 1:2.13 (Hyde et al., 2008; Machin & Pekkarinen, 2008; Reilly et al., 2015; Stoet & Geary, 2013). Importantly, the female-to-male ratio varied substantially across countries.

The Role of Gender Equality for Gender Differences in Top-Performing Math Students

Guided by social role theory and expectancy-value theory, the second goal of this meta-analysis was to investigate the moderating role of gender equality indicators for gender differences in top-performing math students. Both theories predict that gender differences should be smaller in societies that endorse gender-typed roles to a smaller extent and thus

have greater gender equality. Furthermore, differences in students' achievement scores between different (gender-typed) achievement domains should be smaller in countries that have greater gender equality. We were able to explain some of the heterogeneity in effect sizes with domain-specific gender equality indicators as moderators. Our results suggested that tertiary enrollment ratios predicted the proportion of female students in the top 5% in mathematics. That is, the proportion of female students was higher when the share of female students enrolled in tertiary education was higher. Relative to previous studies (Guiso et al., 2008; Hyde & Metz, 2009; Penner, 2008) using composite gender equality indicators (e.g., the Global Gender Gap Index) that aggregate multiple domains of gender equality into one value, our results point to specific domains of gender equality that may be directly or indirectly responsible for the development of mathematical talent in female students. Furthermore, tertiary enrollment ratios and women's share of research positions in a country predicted mathematically top-performing female and male students' profile scores. For female students, math–reading and science–reading profile scores became less pronounced (i.e., the achievement scores in two domains differed less) when the share of female students enrolled in tertiary education was higher. For male students, this relation was found for their math–reading profile scores. Moreover, female students' scores in mathematics and science differed less when the proportion of women in research positions in a country was higher.

However, gender gaps in achievement and achievement motivation in math, reading, and science in the group of top-performing math students were not related to domain-specific gender equality indicators. The most likely explanation for this is that the effect sizes did not vary much within or between countries as indicated by standard deviations of zero or close to zero within countries or at the country level. This restriction of range may have hampered the ability to detect moderating relations between gender equality indicators on the one hand and

gender differences in achievement and achievement motivation in top-performing students on the other.

Overall, our results suggest that in societies that value higher education for women, more female students score in the top 5% in mathematics. In addition, achievement differences in different domains are smaller for female students (and partially also for male students), the more women study at universities and the more women hold research positions. Thus, the (realistic) perspective of attending a university and entering research positions for female students might (a) motivate female students to develop mathematical talent and (b) motivate female and male students to develop skills in several areas at a more similar level. Thus, our results (at least partially) support the predictions of expectancy-value theory (Eccles et al., 1983) and social role theory (Eagly, 1987).

Practical Implications

STEM professions are important to a country's competitiveness and economic well-being (Halpern et al., 2007). Thus, successfully recruiting talented future professionals in this field is one major concern of modern societies. However, women are still underrepresented in STEM careers, especially in the fields of engineering, physics, and computer science. In 2013, women made up only 29% of the science and engineering workforce but accounted for half of the total college-educated workforce in the US (National Science Board, 2016). Hence, making fuller use of the female talent pool could play a vital role in addressing workforce shortages (e.g., Bureau of Labor Statistics, 2019). Furthermore, working in STEM fields also provides positive benefits for women and men: STEM fields usually offer better earning opportunities and better working conditions compared with non-STEM fields (National Science Board, 2016).

In consideration of expectancy-value theory's assumption that gender differences in students' values and expectancies for success are vital factors for later gender differences in

their occupational choices (Eccles, 1994), gender differences in top-performing math students can have implications for women's underrepresentation in STEM fields. Based on the findings from the present meta-analysis, we provide the following presumptions.

The still existing preponderance of male students in the talent pool for STEM careers (i.e., in the right tail of the distribution of math achievement), also found in this meta-analysis, may partly explain women's underrepresentation in STEM. Another potentially contributing factor might be male students' more mathematics-oriented achievement profiles. Having one dominant academic strength is likely to promote higher self-concept in that domain and a clear goal to invest time, effort, and energy into pursuing mathematics-related fields in one's future career. This lines up with our finding that male students reported on average slightly higher math self-concept, self-efficacy, and stronger intentions to choose additional math courses in school and beyond compared with female students. By contrast, having multiple academic strengths is likely to result in more ambiguous expectancies and self-concepts and, consequently, less specific career goals, which is more likely true for mathematically top-performing female students as they had more balanced achievement profiles and stronger verbal motivation than male students (Valla & Ceci, 2014). In other words: "Those who can only do mathematics, do mathematics, but those with multiple extreme talents may choose to do something else" (Ceci et al., 2009, p. S3). Finally, even if mathematically top-performing female students enter STEM careers, given their specific interests in organic sciences, they would be more likely to work in medical fields or biological sciences than in inorganic sciences. By contrast, top-performing male students would be more likely to enter inorganic STEM fields, such as physics or engineering, given their respective science interests.

Strengths, Limitations, and Future Research Directions

The findings of the present meta-analysis represent especially strong scientific evidence because they are based on (a) individual student data (Stewart & Tierney, 2002) from (b)

representative, unselective samples of top-performing math students from well-defined populations, namely 15-year-olds in PISA (Hedges & Nowell, 1995; Reilly et al., 2019). Furthermore, we applied state-of-the-art methods in meta-analyses (i.e., accounting for the dependencies between effect sizes in random-effects models, multiple imputation of moderating variables, multivariate meta-regressions; Tipton et al., 2019). Despite these strengths, the present study has several limitations that should be addressed in future research. First, according to expectancy-value theory, educational and occupational choices are assumed to be influenced by intraindividual hierarchies of achievement and motivation in different domains (Eccles, 1994). Within each PISA cycle, achievement measures in several domains are available, whereas the assessment of achievement motivation was focused on a single domain. Consequently, we were able to analyze gender differences in top-performing math students' achievement profiles but not in their motivational profiles. Gender differences in top-performing math students' motivational profiles should be investigated in future studies. In addition, longitudinal research could investigate whether the interplay between cognitive and motivational profiles predicts top-performing math students' career choices (i.e., STEM vs. non-STEM fields; specific STEM fields).

Another limitation is that PISA employs a cross-sectional design across countries. By using these data, we could not explore the impact of the gender differences found in the present study for top-performing math students' future educational and occupational STEM-related choices. Rather, any predictions we made in the present study were based on theoretical assumptions (Eccles, 1994; Wood & Eagly, 2012) and empirical evidence provided by previous longitudinal research.

Because PISA data are obtained from standardized testing, a further limitation is that we cannot completely rule out that our results are affected by stereotype threat effects. Stereotype threat theory predicts that members of a negatively stereotyped group will

underperform on standardized tests when (a) that stereotype is made salient or relevant for the task at hand, and (b) they are concerned about being judged or treated negatively on the basis of this stereotype (Spencer et al., 2016). To trigger stereotype threat, "simply sitting down to write a test in a negatively stereotyped domain is enough […]." (Spencer et al., 2016, p. 418). Because math and science are stereotypically male domains and reading is stereotypically a female domain, we could not preclude the possibility that stereotype threat impaired female students' performance in math and science or male students' performance in reading (e.g., Hartley & Sutton, 2013; Pansu et al., 2016; Picho et al., 2013). However, if these effects were present in the PISA assessments, they were probably very small because the typical threat scenarios were not activated (e.g., no verbal or written statement that male students are superior to female students on the test, no priming of female identity; e.g., OECD, 1999, 2005a). Instead, PISA is designed in such a way that student achievement is assessed first and then students are asked to indicate their gender on a subsequent student questionnaire (OECD, 2002, 2005b, 2009, 2012, 2013, 2017b). Moreover, it should be noted that recent research on stereotype threat in secondary education has shown divergent findings and that the literature seems to be distorted by publication bias (Flore et al., 2018; Flore & Wicherts, 2015; Shewach et al., 2019; Wei, 2012).

Another limitation is that although the present meta-analysis covered a large number of countries representing about 90% of the world economy (Schleicher, 2007), data were not available for all countries around the world. Participation rates in PISA are especially low in low- and lower-middle-income countries, mainly because participation in PISA is associated with high costs and high demands on the assessment infrastructure of a country (Lockheed et al., 2015). Nevertheless, a more diverse sample of countries would be desirable to draw even more generalizable conclusions. For example, we show in Table S21 that female students have on average less access to formal schooling at the primary, secondary, and tertiary levels

of education in countries that did not participate in PISA than in countries that participated. Thus, it is likely that gender differences would be larger in a more diverse sample.

Conclusions

Capitalizing on representative individual student data for top-performing math students from 82 countries, the present meta-analysis makes four major contributions. First, we showed that, on average, two out of five adolescent students in the top 5% in mathematics are female. Second, we found that mathematically top-performing female and male students were similar with regard to their achievement in mathematics, reading, and science and in most characteristics that are related to achievement motivation in mathematics and science. Third, we provided strong empirical evidence that male students tended to have mathematics-oriented achievement profiles, whereas female students' achievement profile scores were more balanced. Additionally, female students had stronger motivation levels in reading than male students. Furthermore, we found important gender differences in top-performing math students' specific science interests: Whereas male students were more interested in learning about physics- and engineering-related topics, female students expressed greater interest in health- and biology-related domains. Fourth, tertiary enrollment ratios and women's share of research positions were related to the proportion of female students in the top 5% in mathematics and students' achievement profiles. To conclude, the results of the present meta-analysis demonstrate that there are important gender differences in top-performing math-students' achievement profiles and verbal and science motivation at the end of compulsory education.

References

Ahmed, I., Sutton, A. J., & Riley, R. D. (2012). Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: A database survey. *BMJ, 344*(d7762). https://doi.org/10.1136/bmj.d7762

Ashmore, R. D., & Del Boca, F. K. (1979). Sex stereotypes and implicit personality theory: Towards a cognitive–social psychological conceptualization. *Sex Roles, 5*, 219–248. https://doi.org/10.1007/BF00287932

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66*(2), 91–103. https://www.jstor.org/stable/2112795

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education, 4*(1). https://doi.org/10.1186/s40536-015-0015

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431–449. https://dx.doi.org/10.1037/a0038047

Brunner, M., Keller, U., Hornung, C., Reichert, M., & Martin, R. (2009). The cross-cultural generalizability of a new structural model of academic self-concepts. *Learning and Individual Differences, 19*(4), 387–403. https://doi.org/10.1016/j.lindif.2008.11.008

Bureau of Labor Statistics (2019). *Computer and information research scientist.* Retrieved from https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 15*(3), 75–141. https://doi.org/10.1177/1529100614541236

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

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*(1521). https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*(738). https://doi.org/10.3389/fpsyg.2016.00738

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129. https://doi.org/10.2307/3001666

Cohen, J. (1988). *Statistical power analysis for the behavioral science*. Lawrence Erlbaum.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165–176. https://dx.doi.org/10.1037/a0015565

Coyle, T. R., Purcell, J. M., Snyder, A. C., & Richmond, M. C. (2014). Ability tilt on the SAT and ACT predicts specific abilities and college majors. *Intelligence, 46,* 18–24. https://doi.org/10.1016/j.intell.2014.04.008

Coyle, T. R., Snyder, A. C., & Richmond, M. C. (2015). Sex differences in ability tilt: Support for investment theory. *Intelligence, 50,* 209–220. https://doi.org/10.1016/j.intell.2015.04.012

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data.* OECD Publishing. https://dx.doi.org/10.1787/9789264248373-en

Dekhtyar, S., Weber, D., Helgertz, J., & Herlitz, A. (2018). Sex differences in academic strengths contribute to gender segregation in education and occupation: A longitudinal examination of 167,776 individuals. *Intelligence, 67,* 84–92. https://doi.org/10.1016/j.intell.2017.11.007

Diekman, A. B., Brown, E., Johnston, A., & Clark, E. (2010). Seeking congruity between goals and roles: A new look at why women opt out of STEM careers. *Psychological Science, 21*(8), 1051–1057. https://doi.org/10.1177/0956797610377342

Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., & Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to stem careers: Evidence for a goal congruity perspective. *Journal of Personality and Social Psychology, 101*(5), 902–918. https://doi.org/10.1037/a0025199

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Erlbaum.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist 75*(3), 301–315. https://doi.org/10.1037/amp0000494

Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of Women Quarterly, 18*(4), 585–609. https://doi.org/10.1111/j.1471-6402.1994.tb01049.x

Eccles (Parsons), J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). In J. T. Spence (Ed.), *Expectations, values and academic*

behaviors. Perspective on achievement and achievement motivation (pp. 75–146). W. H. Freeman.

Else-Quest, N. M., & Grabe, S. (2012). The political is personal: Measurement and application of nation-level indicators of gender equity in psychological research. *Psychology of Women Quarterly, 36*(2), 131–144. https://doi.org/10.1177/0361684312441592

Else-Quest, N. M., Hyde, J. S., Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Eurostat (n.d.). *Tertiary education statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php/Tertiary_education_statistics#Fields_of_study

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A Registered Report. *Comprehensive Results in Social Psychology, 3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods, 21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological*

Science in the Public Interest, 8(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hartley, B. L., & Sutton, R. M. (2013). A stereotype threat account of boys' academic underachievement. *Child Development, 84*(5), 1716–1733. https://doi.org/10.1111/cdev.12079

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45. https://doi.org/10.1126/science.7604277

Hempel, S., Miles, J. N., Booth, M. J., Wang, Z., Morton, S. C., & Shekelle, P. G. (2013). Risk of bias: A simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Reviews, 2*(1), 107. https://doi.org/10.1186/2046-4053-2-107

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78–80. https://doi.org/10.1037/0003-066X.58.1.78

Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011). The Cochrane Collaboration. Available from www.handbook.cochrane.org.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. https://doi:10.1002/sim.1186

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592. https://dx.doi.org/10.1037/0003-066X.60.6.581

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

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14*(3), 299–324. https://doi.org/10.1111/j.1471-6402.1990.tb00022.x

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

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences, 106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

Jacobs, J., Davis-Kean, P., Bleeker, M., Eccles, J., & Malanchuk, O. (2005). "I can, but I don't want to": The impact of parents, interests, and activities on gender differences in math. In A. Gallagher & J. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 73–98). Cambridge University Press.

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology, 107*(3), 371–392. http://dx.doi.org/10.1037/a0037215

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Sage.

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). *The experience of middle-income countries participating in PISA 2000-2015.* OECD Publishing/World Bank. https://dx.doi.org/10.1787/9789264246195-en

Lohman, D. F., Gambrell, J., & Lakin, J. (2008). The commonality of extreme discrepancies in the ability profiles of academically gifted students. *Psychology Science, 50*(2), 269–282.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*(4), 718–729. https://doi.org/1037//0021-9010.86.4.718

Lytton, H., & Romney, D. M. (1991). Parents' differential socialization of boys and girls: A meta-analysis. *Psychological Bulletin, 109*(2), 267–296. https://doi.org/10.1037/0033-2909.109.2.267

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 322*(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Makel, M. C., Wai, J., Peairs, K., & Putallaz, M. (2016). Sex differences in the right tail of cognitive abilities: An update and cross cultural extension. *Intelligence, 59,* 8–15. https://doi.org/10.1016/j.intell.2016.09.003

McDaniel, A. (2016). The role of cultural contexts in explaining cross-national gender gaps in STEM expectations. *European Sociological Review, 32*(1), 122–133. https://doi.org/10.1093/esr/jcv078

Meece, J. L., Eccles-Parsons, J., Kaczala, C. M., Goff, S. B., & Futterman, R. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin, 91,* 324–348. https://dx.doi.org/10.1037/0033-2909.91.2.324

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology, 107*(3), 631–644. https://doi.org/10.1037/edu0000005

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences, 18*(1), 37–45. https://dx.doi.org/10.1016/j.tics.2013.10.011

Morris, M. L. (2016). Vocational interests in the United States: Sex, age, ethnicity, and year effects. *Journal of Counseling Psychology, 63*(5), 604–615. https://dx.doi.org/10.1037/cou0000164

Naemi, B., Gonzalez, E., Bertling, J., Betancourt, A., Burrus, J., Kyllonen, P. C., Minsky, J., Lietz, P., Klieme, E., Vieluf, S., Lee, J., & Roberts, R. D. (2013). Large-scale group score assessments: Past, present, and future. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean, *Oxford handbook of child psychological assessment* (pp. 129–149). Oxford University Press.

National Science Board (2016). *Science and engineering indicators 2016*. National Science Foundation.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Yoav Bar-Anana, Y., Berghd, R., Caie, H., Gonsalkoralef, K., Kesebira, S., Maliszewskig, N., Netoh, F., Ollii, E., Parkj, J., Schnabelk, K., Shiomural, K., Tulburem, B. T., Wiersn, R. W., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39*(1–2), 21–43. https://doi.org/10.1023/A:1018873615316

OECD (1999). *Test administrator's manual.* https://www.acer.org/files/pisa2000_test_administrator_manual.pdf

OECD (2002). *PISA 2000 technical report.* OECD Publishing.

OECD (2003). *Learners for life: Student approaches to learning–Results from PISA 2000.* OECD Publishing.

OECD (2005a). *Test administrator's manual.* https://www.acer.org/files/pisa2006_test_administrator_manual.pdf

OECD (2005b). *PISA 2003 technical report*. OECD Publishing.

OECD (2009). *PISA 2006 Technical report.* OECD Publishing.

OECD (2010). *PISA 2009 Results: Learning to learn – Student engagement, strategies and practices* (Vol. III). OECD Publishing. https://dx.doi.org/10.1787/9789264083943-en

OECD (2012). *PISA 2009 Technical report.* OECD Publishing.

OECD (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. https://dx.doi.org/10.1787/9789264190511-en

OECD (2017a). *PISA 2015 Technical report.* OECD Publishing.

OECD (2017b). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving* (revised ed.). OECD Publishing. http://dx.doi.org/10.1787/9789264281820-en

Olszewski-Kubilius, P., & Lee, S. Y. (2011). Gender and other group differences in performance on off-level tests: Changes in the 21st century. *Gifted Child Quarterly, 55*(1), 54–73. https://doi.org/10.1177/0016986210382574

Pansu, P., Régner, I., Max, S., Colé, P., Nezlek, J. B., & Huguet, P. (2016). A burden for the boys: Evidence of stereotype threat in boys' reading performance. *Journal of Experimental Social Psychology, 65*, 26–30. https://dx.doi.org/10.1016/j.jesp.2016.02.008

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the arts and sciences. *Psychological Science, 18*(11), 948–952. https://doi.org/10.1111/j.1467-9280.2007.02007.x

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology, 114*(S1), 138–170. https://doi.org/10.1016/j.ssresearch.2007.06.012

Picho , K., Rodriguez, A. & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A meta-analysis. *The Journal of Social Psychology, 153*(3), 299–333. https://doi.org/10.1080/00224545.2012.737380

Pigott, T. D. (2019). Missing data in meta-analysis (pp. 367–382). In H. Cooper, L. V. Hedges, & J. C. Valentine, *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

R Core Team (2019). *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna). Retrieved from: https://www.R-project.org/

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PlOS ONE, 7*(7), e39904. https://doi.org/10.1371/journal.pone.0039904

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

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445–458. https://dx.doi.org/10.1037/amp0000356

Riegle-Crumb, C. (2005). The cross-national context of the gender gap in math and science. In L. Hedges & B. Schneider (Eds.), *The social organization of schooling* (pp. 227–243). Russell Sage Foundation.

Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change, 8*(4), 349–357. https://doi.org/10.1007/s10833-007-9042-x

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology, 104*(12), 1514–1534. https://dx.doi.org/10.1037/apl0000420

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science, 1,* 115–120. https://doi.org/10.1177/2515245918757689

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67,* 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & The Health Professions, 25*(1), 76–97. https://dx.doi.org/10.1177/0163278702025001006

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PLOS ONE, 8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence, 48,* 137–151. https://doi.org/10.1016/j.intell.2014.11.006

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6*(189). https://doi.org/10.3389/fpsyg.2015.00189

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

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods, 10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region.* ACER Publishing.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215–247. https://doi.org/10.3102/1076998609346961

Valla, J. M., & Ceci, S. J. (2014). Breadth-based models of women's underrepresentation in STEM fields: An integrative commentary on Schmidt (2011) and Nye et al. (2012). *Perspectives on Psychological Science, 9*(2), 219–224. https://doi.org/10.1177/1745691614522067

van Erp, S., Verhagen, J., Grasman, R.P.P.P. and Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin

from 1990–2013. *Journal of Open Psychology Data, 5*(1), 4. https://doi.org/10.5334/jopd.33

Wai, J., Hodges, J., & Makel, M. C. (2018). Sex differences in ability tilt in the right tail of cognitive abilities: A 35-year examination. *Intelligence, 67,* 76–83. https://doi.org/10.1016/j.intell.2018.02.003

Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340.

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

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, M. T., Ye, F., & Degol, J. L. (2017). Who chooses STEM careers? Using a relative cognitive strength and interest model to predict careers in science, technology, engineering, and mathematics. *Journal of Youth and Adolescence, 46*(8), 1805–1820. https://doi.org/10.1007/s10964-016-0618-8

Wei, T. E. (2012). Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation and Policy Analysis, 34*(4), 465–488. https://doi.org/10.3102/0162373712452629

Wilgenbusch, T., & Merrell, K. W. (1999). Gender differences in self-concept among children and adolescents: A meta-analysis of multidimensional studies. *School Psychology Quarterly, 14*(2), 101–120. https://dx.doi.org/10.1037/h0089000

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In J. M. Olson & M. P. Zanna, *Advances in experimental social psychology* (Vol. 46, pp. 55–123). Academic Press.

Table 1. *Number of Participating Countries, Sample Sizes, and Mean Percentages of Female Students (%F) in the Full PISA Sample and in the Sample used in the Present Study (Top-Performing Math Students) per Cycle and in Total*

	Full PISA sample			Top 5% in mathematics		
PISA	Countries	*N*	%F	Countries	*N*	%F
2000	43	127,388	50	42	6,314	39
2003	41	276,165	50	40	13,752	37
2006	57	398,750	51	56	19,920	39
2009	73	515,958	51	72	25,781	40
2012	65	480,174	50	64	23,994	39
2015	69	482,067	50	69	25,720	41
Total	83	2,280,502	50	82	115,481	40

Note. PISA = Programme for International Student Assessment. In PISA 2015, data from Argentina, Malaysia, and Kazakhstan were not included because either their population or construct were inadequately covered (OECD, 2017a). Spain additionally assessed its 17 adjudicated regions in PISA 2015. These data were not included in the present study or in this table. Furthermore, the US additionally assessed a subsample of federal states (in PISA 2012 and 2015) and Puerto Rico (PISA 2015), which were not analyzed due to data policies designed to protect the confidentiality of individually identifiable information.

Table 2. *Indicators of Gender Equality Used in the Present Study*

Indicator	Description	Cycles
Women's share of higher positions[a,b]	Women's share of employment in senior and middle management (%), i.e., in decision-making and management roles in government, large enterprises and institutions	ILO: 2006, 2009, 2012, 2015 UN: 2000, 2012, 2015
Women's share of research positions[c,d]	Percentage of research positions held by women	OECD: 2000, 2003, 2006, 2009, 2012, 2015 UNESCO: 2000, 2003, 2006, 2009, 2012, 2015
Primary enrollment ratio[c]	Ratio of the percentages of female (numerator) and male students (denominator) in the population of official school-age students enrolled in primary education	2000, 2003, 2006, 2009, 2012, 2015
Secondary enrollment ratio[c]	Ratio of the percentages of female (numerator) and male students (denominator) in the population of official school-age students enrolled in secondary education	2000, 2003, 2006, 2009, 2012, 2015
Tertiary enrollment ratio[c]	Ratio of the percentages of female (numerator) and male students (denominator) in the population of official school-age students enrolled in tertiary education	2000, 2003, 2006, 2009, 2012, 2015

Note. For all indicators, higher values indicate greater gender equality; ILO = International Labour Organization; UN = United Nations; OECD = Organisation for Economic Co-operation and Development.
[a]Available from https://hdr.undp.org/
[b]Available from https://ilostat.ilo.org/data/
[c]Available from https://data.uis.unesco.org
[d]Available from https://stats.oecd.org

Table 3. *Meta-Analytic Results on Gender Differences in Achievement and Achievement Motivation in Top-Performing Math Students (Top 5%)*

Outcome	ES	$Mean_w$	95% CI	N_{CNT}	k	Q	T_{total}	T_{Level2}	T_{Level3}	I^2_{total}	I^2_{Level2}	I^2_{Level3}	MA
Percentage of female students in the top 5% in mathematics	%	40.10	[38.93, 41.27]	82	343	5402.80***	5.90	3.24	4.93	95	29	66	✓
Gender differences in achievement													
Math	*d*	0.05	[0.03, 0.06]	82	343	29.41	0.00	0.00	0.00	0	0	0	
Reading	*d*	-0.23	[-0.25, -0.21]	82	342	610.64***	0.12	0.12	0.00	46†	46	0	✓
Science	*d*	0.01	[-0.01, 0.02]	82	343	104.65	0.00	0.00	0.00	0	50	50	✓
Gender differences in math motivation													
Self-efficacy	*d*	0.32	[0.28, 0.35]	65	104	159.18***	0.10	0.00	0.10	39†	0	39	✓
Intention	*d*	0.27	[0.23, 0.31]	64	64	96.11**	0.09	–	–	35†	–	–	✓
Instrumental motivation	*d*	0.16	[0.13, 0.19]	66	137	164.40*	0.08	0.00	0.08	22	0	22	✓
Self-concept	*d*	0.15	[0.12, 0.18]	66	137	175.72*	0.08	0.00	0.08	28	0	28	✓
Interest	*d*	0.10	[0.07, 0.13]	66	137	163.21	0.08	0.00	0.08	24	0	24	
Subjective norms	*d*	0.01	[-0.03, 0.04]	64	64	89.00*	0.08	–	–	29	–	–	✓
Attribution of failure	*d*	-0.13	[-0.16, -0.09]	64	64	66.33	0.05	–	–	17	–	–	
Anxiety	*d*	-0.15	[-0.18, -0.12]	65	104	141.87**	0.09	0.03	0.08	34†	4	31	✓
Work ethic	*d*	-0.16	[-0.20, -0.12]	64	64	80.00	0.08	–	–	26	–	–	
Gender differences in reading motivation													
Verbal self-concept	*d*	-0.38	[-0.46, -0.30]	33	33	35.49	0.09	–	–	14	–	–	
Interest	*d*	-0.50	[-0.57, -0.44]	33	33	29.93	0.02	–	–	1	–	–	
Enjoyment	*d*	-0.64	[-0.69, -0.60]	73	114	248.75***	0.15†	0.00	0.15	58†	0	58	✓
Gender differences in science motivation													
Self-concept	*d*	0.19	[0.14, 0.23]	56	56	110.00***	0.11	–	–	50†	–	–	✓
General value	*d*	0.12	[0.08, 0.16]	56	56	75.13*	0.08	–	–	29	15	15	✓
Self-efficacy	*d*	0.07	[0.04, 0.11]	72	124	232.95***	0.12†	0.00	0.12	51†	0	51	✓
Future-oriented motivation	*d*	0.05	[0.00, 0.11]	56	56	111.08***	0.13†	–	–	50†	–	–	✓
Enjoyment	*d*	0.07	[0.03, 0.10]	72	124	193.50***	0.10	0.00	0.10	43†	0	43	✓

(table continues)

Table 3 *(Continued)*

Outcome	ES	$Mean_w$	95% CI	N_{CNT}	k	Q	T_{total}	T_{Level2}	T_{Level3}	I^2_{total}	I^2_{Level2}	I^2_{Level3}	MA
Gender differences in science motivation													
Personal value	*d*	0.05	[0.01, 0.08]	56	56	62.45	0.05	–	–	14	–	–	
Instrumental motivation	*d*	0.00	[-0.02, 0.03]	72	124	155.36*	0.07	0.04	0.06	26	7	20	✓
Gender differences in science interest													
Motion and forces	*d*	0.54	[0.50, 0.58]	55	55	69.06	0.07	–	–	22	–	–	
Physics	*d*	0.40	[0.35, 0.44]	56	56	88.81**	0.11	–	–	37†	–	–	✓
Energy transformation	*d*	0.39	[0.36, 0.43]	55	55	65.66	0.05	–	–	15	–	–	
History of the universe	*d*	0.14	[0.10, 0.17]	55	55	47.36	0.00	–	–	0	–	–	
Chemistry	*d*	0.06	[0.02, 0.10]	56	56	67.02	0.06	–	–	17	–	–	
Geology	*d*	0.04	[0.01, 0.08]	56	56	32.44	0.00	–	–	0	–	–	
Astronomy	*d*	-0.04	[-0.08, -0.01]	56	56	40.81	0.00	–	–	0	–	–	
Biosphere	*d*	-0.11	[-0.15, -0.08]	55	55	38.03	0.00	–	–	0	–	–	
Plant biology	*d*	-0.30	[-0.34, -0.25]	56	56	55.45	0.06	–	–	13	–	–	
Disease	*d*	-0.30	[-0.34, -0.27]	55	55	54.19	0.01	–	–	1	–	–	
Human biology	*d*	-0.44	[-0.48, -0.40]	56	56	69.15	0.08	–	–	26	–	–	

Note. ES = Type of effect size (percentage or Cohen's *d*); $Mean_w$ = Weighted mean effect size; 95% CI = 95% confidence interval; N_{CNT} = Number of countries; k = number of effect sizes, Q = Total homogeneity statistic; T_{Level2} = Within-countries standard deviation of effect sizes, T_{Level3} = Between-countries standard deviation of effect sizes; I^2_{Level2} = Percentage of the variability in effect sizes that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in effect sizes that is due to heterogeneity between countries rather than sampling error. A dash in the columns T_{Level2}/T_{Level3} and I^2_{Level2} /I^2_{Level3} indicates that a two-level random effects model was used to analyze a single PISA cycle. MA = Was a moderator analysis conducted (based on heterogeneity measures)? ✓ = Yes.

* $p < .05$, ** $p < .01$, *** $p < .001$

† Moderate heterogeneity (i.e., $T \geq 0.12$ for standardized mean differences or $I^2 \geq 30\%$)

Table 4. *Meta-Regression Models for Explaining Heterogeneity in Gender Differences in the Group of Top-Performing Math Students (Top 5%)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2	R^2_{L2}	R^2_{L3}
Percent female students in the top 5% in mathematics	hp	0.117	[-0.302, 0.536]	.584	82	343	5402.8	–	4	12
	wr	-0.060	[-0.501, 0.380]	.788						
	per	0.034	[-0.084, 0.153]	.570						
	ser	0.011	[-0.068, 0.090]	.789						
	ter	**0.041**	[0.009, 0.073]	.012						
Reading achievement	hp	0.001	[-0.059, 0.062]	.964	82	342	610.64	–	2	0
	wr	-0.001	[-0.075, 0.072]	.969						
	per	0.000	[-0.011, 0.011]	.997						
	ser	0.000	[-0.007, 0.007]	.953						
	ter	0.000	[-0.004, 0.004]	.989						
Science achievement	hp	0.000	[-0.054, 0.053]	.990	82	343	104.65	–	99	95
	wr	0.000	[-0.062, 0.062]	1						
	per	0.000	[-0.009, 0.009]	.985						
	ser	0.000	[-0.006, 0.006]	.907						
	ter	0.000	[-0.003, 0.003]	.981						
Math self-efficacy	hp	-0.004	[-0.082, 0.073]	.915	65	104	159.18	–	0	25
	wr	0.003	[-0.093, 0.099]	.952						
	per	0.001	[-0.014, 0.015]	.930						
	ser	-0.001	[-0.010, 0.009]	.910						
	ter	0.000	[-0.005, 0.004]	.921						
Math intention	hp	0.000	[-0.087, 0.086]	.992	64	64	96.11	34	–	–
	wr	0.005	[-0.102, 0.112]	.926						
	per	-0.002	[-0.024, 0.021]	.872						
	ser	-0.002	[-0.013, 0.010]	.792						
	ter	0.000	[-0.005, 0.005]	.995						
Instrumental math motivation	hp	-0.001	[-0.077, 0.075]	.973	66	137	164.40	–	0	14
	wr	0.001	[-0.091, 0.093]	.982						
	per	0.001	[-0.014, 0.016]	.867						
	ser	0.001	[-0.008, 0.009]	.898						
	ter	0.000	[-0.005, 0.004]	.866						
Math self-concept	hp	-0.004	[-0.077, 0.069]	.918	66	137	175.72	–	0	44
	wr	0.001	[-0.086, 0.089]	.977						
	per	0.002	[-0.013, 0.017]	.821						
	ser	0.001	[-0.008, 0.009]	.908						
	ter	0.000	[-0.005, 0.004]	.897						

(table continues)

Table 4 *(Continued)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2	R^2_{L2}	R^2_{L3}
Subjective math norms	hp	-0.003	[-0.085, 0.080]	.952	64	64	89.00	–	–	17
	wr	0.001	[-0.104, 0.107]	.981						
	per	-0.006	[-0.028, 0.017]	.630						
	ser	-0.001	[-0.012, 0.010]	.878						
	ter	0.000	[-0.005, 0.005]	.966						
Math anxiety	hp	0.003	[-0.074, 0.079]	.944	65	104	141.87	–	99	26
	wr	-0.004	[-0.098, 0.089]	.925						
	per	-0.002	[-0.016, 0.012]	.789						
	ser	-0.002	[-0.011, 0.007]	.717						
	ter	0.001	[-0.004, 0.006]	.791						
Enjoyment of reading	hp	0.001	[-0.080, 0.082]	.981	73	114	248.75	–	0	43
	wr	-0.007	[-0.105, 0.090]	.880						
	per	0.002	[-0.016, 0.019]	.861						
	ser	0.002	[-0.009, 0.013]	.720						
	ter	-0.002	[-0.007, 0.003]	.455						
Science self-concept	hp	-0.008	[-0.105, 0.089]	.873	56	56	110.00	–	–	54
	wr	0.006	[-0.099, 0.112]	.908						
	per	0.001	[-0.022, 0.024]	.919						
	ser	-0.002	[-0.013, 0.010]	.758						
	ter	-0.001	[-0.007, 0.005]	.711						
General value of science	hp	-0.004	[-0.105, 0.096]	.936	56	56	75.13	–	–	31
	wr	0.004	[-0.102, 0.110]	.943						
	per	0.000	[-0.021, 0.020]	.983						
	ser	-0.003	[-0.015, 0.008]	.578						
	ter	0.000	[-0.006, 0.007]	.908						
Science self-efficacy	hp	-0.004	[-0.090, 0.082]	.928	72	124	232.95	–	0	21
	wr	0.004	[-0.093, 0.100]	.940						
	per	0.001	[-0.013, 0.015]	.896						
	ser	-0.001	[-0.010, 0.009]	.881						
	ter	0.001	[-0.005, 0.006]	.834						
Future-oriented motivation	hp	-0.007	[-0.112, 0.097]	.892	56	56	111.08	–	–	46
	wr	0.004	[-0.111, 0.119]	.946						
	per	0.007	[-0.017, 0.031]	.556						
	ser	-0.005	[-0,018, 0.008]	.437						
	ter	0.000	[-0.007, 0.007]	.977						

(table continues)

Table 4 *(Continued)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2	R^2_{L2}	R^2_{L3}
Enjoyment of science	hp	-0.007	[-0.084, 0.071]	.869	72	124	193.50	–	0	50
	wr	0.002	[-0.085, 0.090]	.959						
	per	0.003	[-0.010, 0.017]	.624						
	ser	-0.001	[-0.010, 0.008]	.833						
	ter	0.000	[-0.005, 0.004]	.880						
Instrumental science motivation	hp	-0.005	[-0.078, 0.068]	.896	72	124	155.36	–	70	90
	wr	0.003	[-0.081, 0.086]	.952						
	per	0.001	[-0.012, 0.015]	.862						
	ser	-0.002	[-0.010, 0.007]	.677						
	ter	0.000	[-0.004, 0.004]	.956						
Interest in physics	hp	-0.006	[-0.107, 0.094]	.900	56	56	88.81	94	–	–
	wr	0.011	[-0.100, 0.121]	.851						
	per	0.004	[-0.021, 0.030]	.753						
	ser	-0.005	[-0.017, 0.008]	.474						
	ter	0.000	[-0.006, 0.007]	.939						

Note. N_{CNT} = Number of countries; R^2 = Variance explained in %; R^2_{L2} = Variance explained within countries in %; R^2_{L3} = Variance explained between countries in %; hp = Women's share of higher positions (i.e., legislators, senior officials, managers) in percent; rp = Women's share of research positions in percent; per = Log-transformed ratio of female to male students enrolled in primary education; ser = Log-transformed ratio of female to male students enrolled in secondary education; ter = Log-transformed ratio of female to male students enrolled in tertiary education; Bold values indicate significant results ($p < .05$).

Table 5. *Meta-Analytic Results on Achievement Profiles in the Group of Top-Performing Math Students (Top 5%)*

Domain	ES	95% CI	N_{CNT}	k	Q	T_{total}	T_{Level2}	T_{Level3}	I^2_{total}	I^2_{Level2}	I^2_{Level3}	MA
Mean differences in individual profile scores among male and female students												
Male students												
Math-reading tilt	57.65	[52.09, 63.22]	82	342	7040.22***	28.29	15.74	23.51	96†	30	66	✓
Science-reading tilt	32.20	[29.05, 35.36]	82	342	4353.32***	19.31	15.68	11.26	93†	61	32	✓
Math-science tilt	24.55	[20.32, 28.78]	82	343	4853.94***	21.19	11.62	17.72	93†	28	65	✓
Female students												
Math-reading tilt	22.71	[16.61, 28.81]	82	342	5092.31***	30.24	15.35	26.06	95†	24	70	✓
Science-reading tilt	2.08	[-1.02, 5.18]	82	342	2783.15***	18.12	14.12	11.35	89†	54	35	✓
Math-science tilt	19.87	[15.10, 24.65]	82	343	4268.07***	22.91	10.27	20.48	92†	18	73	✓
Nonoverlap between male and female students' distributions of individual profile scores (in %)												
Math-reading profile	44	[42, 45]	82	342	654.69***	7.04	5.53	4.35	50†	31	19	✓
Science-reading profile	42	[40, 44]	82	342	505.73***	6.41	2.49	5.91	43†	6	36	✓
Math-science profile	18	[17, 19]	82	343	204.46	2.19	0.00	2.19	8	0	8	
Percentages of male and female students demonstrating a certain tilt in their individual profile scores (in %)												
Male students												
Math–reading profile: math $tilt_{M–R}$	87	[85, 89]	82	342	5722.47***	1.01	0.63	0.78	96†	38	58	✓
Science–reading profile: science $tilt_{S–R}$	76	[74, 79]	82	342	5395.50***	0.78	0.58	0.52	95†	53	43	✓
Math–science profile: math $tilt_{M–S}$	69	[66, 72]	82	343	5616.37***	0.75	0.48	0.58	96†	38	58	✓
Female students												
Math–reading profile: math $tilt_{M–R}$	66	[62, 70]	82	342	4785.42***	0.96	0.54	0.80	96†	30	66	✓
Science–reading profile: science $tilt_{S–R}$	52	[49, 55]	82	342	3631.24***	0.66	0.50	0.43	92†	53	39	✓
Math–science profile: math $tilt_{M–S}$	65	[62, 68]	82	343	3987.47***	0.75	0.43	0.62	94†	30	63	✓

Note. ES = Effect size; N_{CNT} = Number of countries; Q = Total homogeneity statistic; T_{Level2} = Within-countries *SD* of effect sizes; T_{Level3} = Between-countries *SD* of effect sizes; I^2_{Level2} = Percentage of the variability in effect sizes that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in effect sizes that is due to heterogeneity between countries rather than sampling error. MA = Was a moderator analysis conducted?;† Moderate heterogeneity (i.e., $I^2 \geq 30\%$).

*** $p < .001$

Table 6. *Meta-Regression Models for Explaining Heterogeneity in Gender Differences in Achievement Profiles in the Group of Top-Performing Math Students (Top 5%)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2_{L2}	R^2_{L3}
			Achievement profile scores						
Math–reading	hp	-0.539	[-1.495, 0.417]	.270	82	342	5092.31	7	22
profile score	wr	-0.623	[-1.714, 0.469]	.263					
female students	per	-0.375	[-1.167, 0.417]	.356					
	ser	-0.033	[-0.386, 0.320]	.856					
	ter	**-0.153**	[-0.271, -0.034]	.012					
Math–reading	hp	-0.642	[-1.569, 0.286]	.175	82	342	7040.22	7	18
profile score	wr	-0.435	[-1.478, 0.608]	.414					
male students	per	-0.331	[-1.138, 0.475]	.423					
	ser	-0.019	[-0.338, 0.299]	.906					
	ter	**-0.123**	[-0.239, -0.006]	.041					
Science–reading	hp	-0.285	[-1.020, 0.451]	.448	82	342	2783.15	3	12
profile score	wr	0.278	[-0.605, 1.160]	.538					
female students	per	-0.048	[-0.591, 0.495]	.863					
	ser	-0.211	[-0.591, 0.169]	.280					
	ter	**-0.097**	[-0.182, -0.012]	.027					
Science–reading	hp	-0.378	[-1.126, 0.370]	.322	82	342	4353.32	3	16
profile score	wr	0.407	[-0.472, 1.285]	.364					
male students	per	0.050	[-0.622, 0.723]	.884					
	ser	-0.250	[-0.626, 0.125]	.194					
	ter	-0.050	[-0.140, 0.041]	.282					
Math–science	hp	-0.208	[-1.092, 0.676]	.645	82	343	4268.07	14	14
profile score	wr	**-0.984**	[-1.904, -0.064]	.036					
female students	per	-0.200	[-0.712, 0.312]	.446					
	ser	0.136	[-0.359, 0.630]	.592					
	ter	-0.013	[-0.121, 0.094]	.806					
Math–science	hp	-0.186	[-1.021, 0.648]	.662	82	343	4853.94	10	20
profile score	wr	-0.828	[-1.740, 0.083]	.075					
male students	per	-0.248	[-0.778, 0.281]	.360					
	ser	0.160	[-0.289, 0.609]	.486					
	ter	-0.071	[-0.175, 0.034]	.187					

(table continues)

Table 6 *(Continued)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2_{L2}	R^2_{L3}
Nonoverlap between male and female students' distributions of individual profile scores									
Nonoverlap	hp	0.001	[-0.050, 0.053]	.955	82	342	654.69	08	20
math–reading	wr	-0.003	[-0.062, 0.056]	.918					
profile	per	0.000	[-0.009, 0.008]	.918					
	ser	0.000	[-0.006, 0.006]	.910					
	ter	0.000	[-0.003, 0.003]	.923					
Nonoverlap	hp	0.002	[-0.051, 0.055]	.948	82	342	505.73	6	21
science–reading	wr	-0.002	[-0.062, 0.057]	.936					
profile	per	-0.001	[-0.009, 0.008]	.905					
	ser	0.000	[-0.006, 0.006]	.965					
	ter	0.000	[-0.004, 0.003]	.791					
Nonoverlap	hp	0.001	[-0,052, 0,053]	.976	82	343	204.46	0	98
math–science	wr	0.000	[-0,055, 0,056]	.986					
profile	per	0.000	[-0,007, 0,008]	.951					
	ser	0.000	]-0,005, 0,005]	.983					
	ter	0.000	[-0,003, 0,003]	.973					
% of male and female students demonstrating a certain tilt in their individual profile scores									
Math–reading	hp	-0.015	[-0.190, 0.159]	.862	82	342	4785.42	4	18
profile math tilt	wr	-0.019	[-0.192, 0.155]	.832					
female students	per	-0.009	[-0.043, 0.024]	.583					
	ser	-0.002	[-0.022, 0.019]	.861					
	ter	-0.004	[-0.015, 0.007]	.440					
Math–reading	hp	-0.025	[-0.204, 0.153]	.780	82	342	5722.47	3	15
profile math tilt	wr	-0.004	[-0.188, 0.180]	.968					
male students	per	-0.008	[-0.046, 0.030]	.681					
	ser	-0.002	[-0.024, 0.020]	.853					
	ter	-0.004	[-0.015, 0.008]	.510					
Science–reading	hp	-0.011	[-0.147, 0.125]	.879	82	342	3631.24	3	9
profile science	wr	0.009	[-0.145, 0.163]	.906					
tilt female	per	-0.002	[-0.029, 0.024]	.856					
students	ser	-0.008	[-0.028, 0.013]	.446					
	ter	-0.002	[-0.012, 0.007]	.598					
Science–reading	hp	-0.017	[-0.163, 0.128]	.814	82	342	5395.50	5	17
profile science	wr	0.024	[-0.139, 0.186]	.775					
tilt male students	per	0.002	[-0.031, 0.034]	.911					
	ser	-0.011	[-0.032, 0.010]	.314					
	ter	0.000	[-0.010, 0.009]	.925					

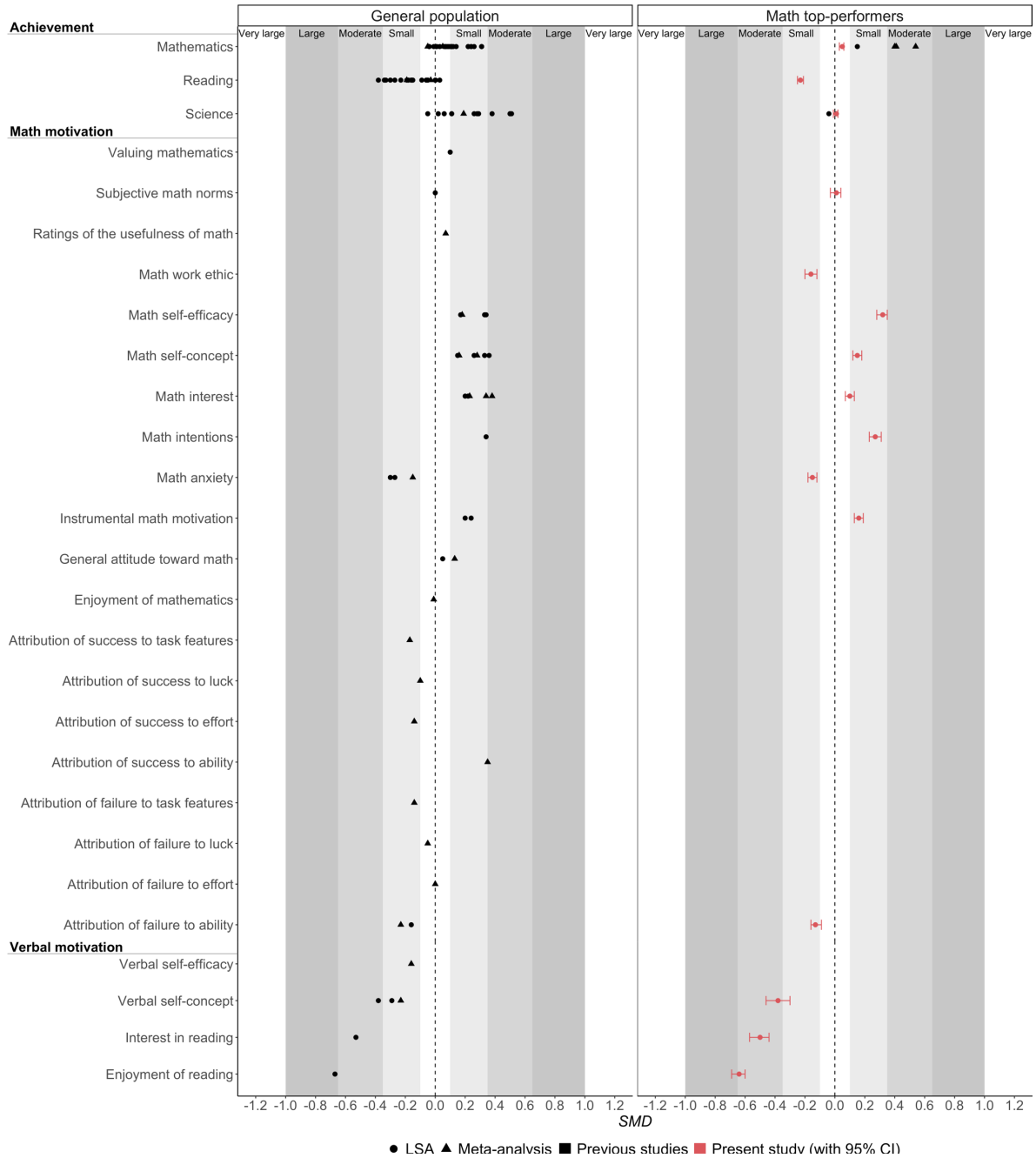
(table continues)

Table 6 *(Continued)*

Outcome	Moderator	*b*	95% CI	*p*	N_{CNT}	*k*	*Q*	R^2_{L2}	R^2_{L3}
% of male and female students demonstrating a certain tilt in their individual profile scores									
Math–science	hp	-0.008	[-0.164, 0.148]	.920	82	343	3987.47	11	10
profile math tilt	wr	-0.030	[-0.187, 0.128]	.713					
female students	per	-0.005	[-0.031, 0.022]	.730					
	ser	0.003	[-0.015, 0.022]	.725					
	ter	-0.001	[-0.011, 0.009]	.819					
Math–science	hp	-0.186	[-1.021, 0.648]	.662	82	343	4853.94	10	20
profile score	wr	-0.828	[-1.740, 0.083]	.075					
male students	per	-0.248	[-0.778, 0.281]	.360					
	ser	0.160	[-0.289, 0.609]	.486					
	ter	-0.071	[-0.175, 0.034]	.187					

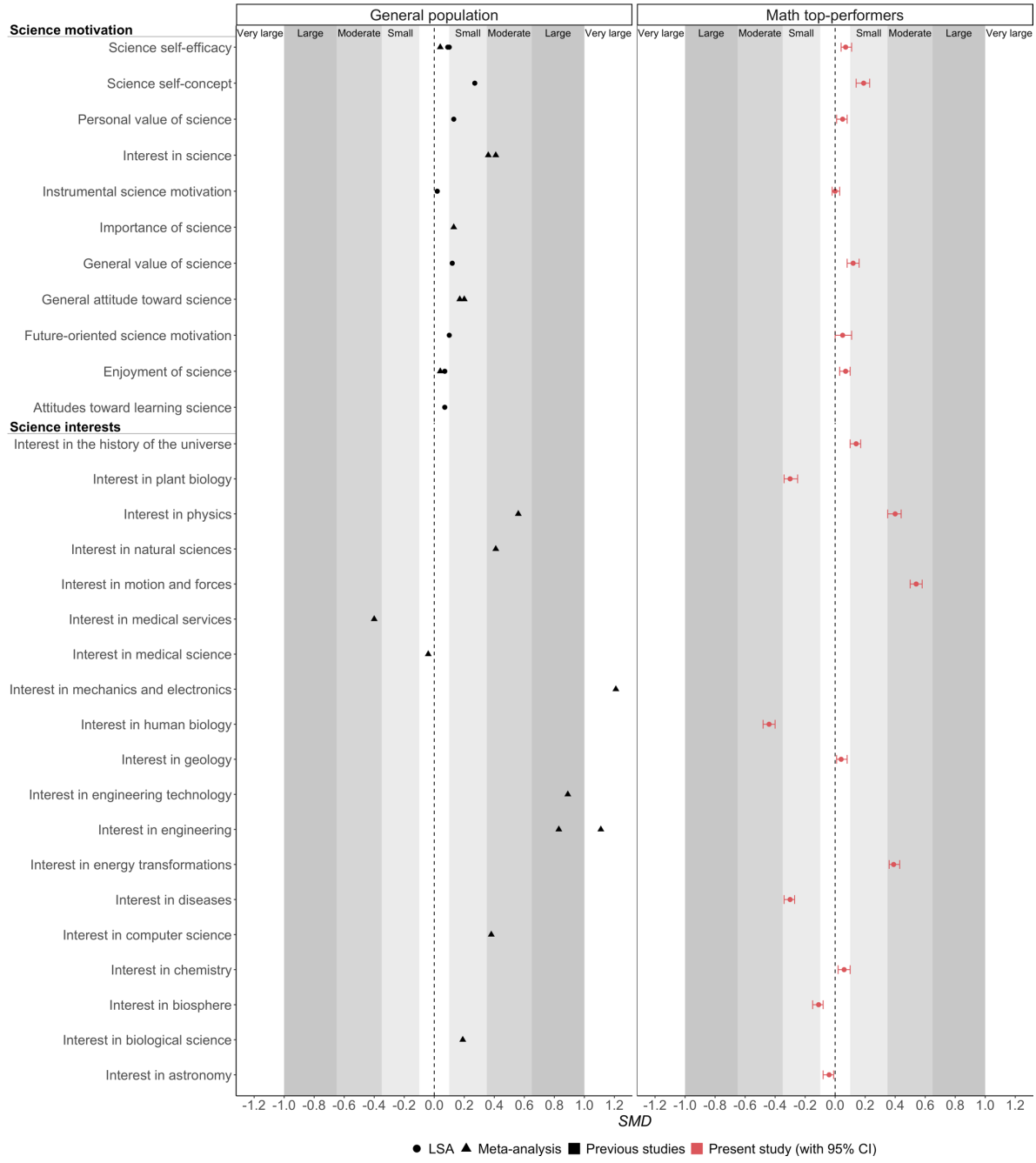
Note. N_{CNT} = Number of countries; R^2_{L2} = Variance explained within countries in %; R^2_{L3} = Variance explained between countries in %; hp = Women's share of higher positions (i.e., legislators, senior officials, managers) in percent; rp = Women's share of research positions in percent; per = Log-transformed ratio of female to male students enrolled in primary education; ser = Log-transformed ratio of female to male students enrolled in secondary education; ter = Log-transformed ratio of female to male students enrolled in tertiary education; Bold values indicate significant results ($p < .05$).

Figure 1. *Overview of Meta-Analyses and Large-Scale Assessments (LSA) on Gender Differences in Math, Reading, and Science Achievement and Math and Verbal Achievement Motivation*



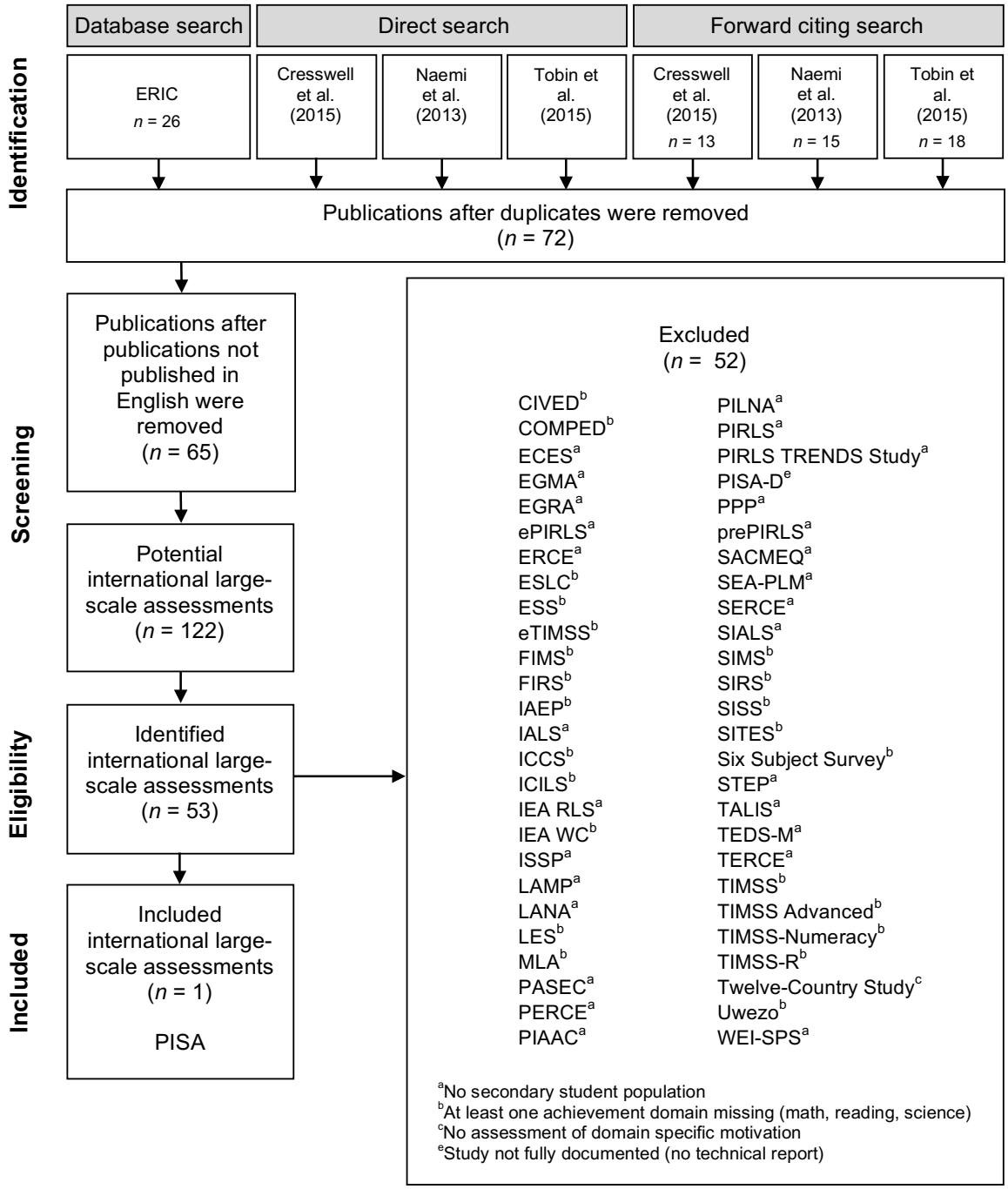
Note. Negative values indicate an advantage of female students, positive values an advantage of male students. See Tables S1 and S2 for more details on all included studies and results on gender ratios.

Figure 2. *Overview of Meta-Analyses and Large-Scale Assessments (LSA) on Gender Differences in Science Achievement Motivation*



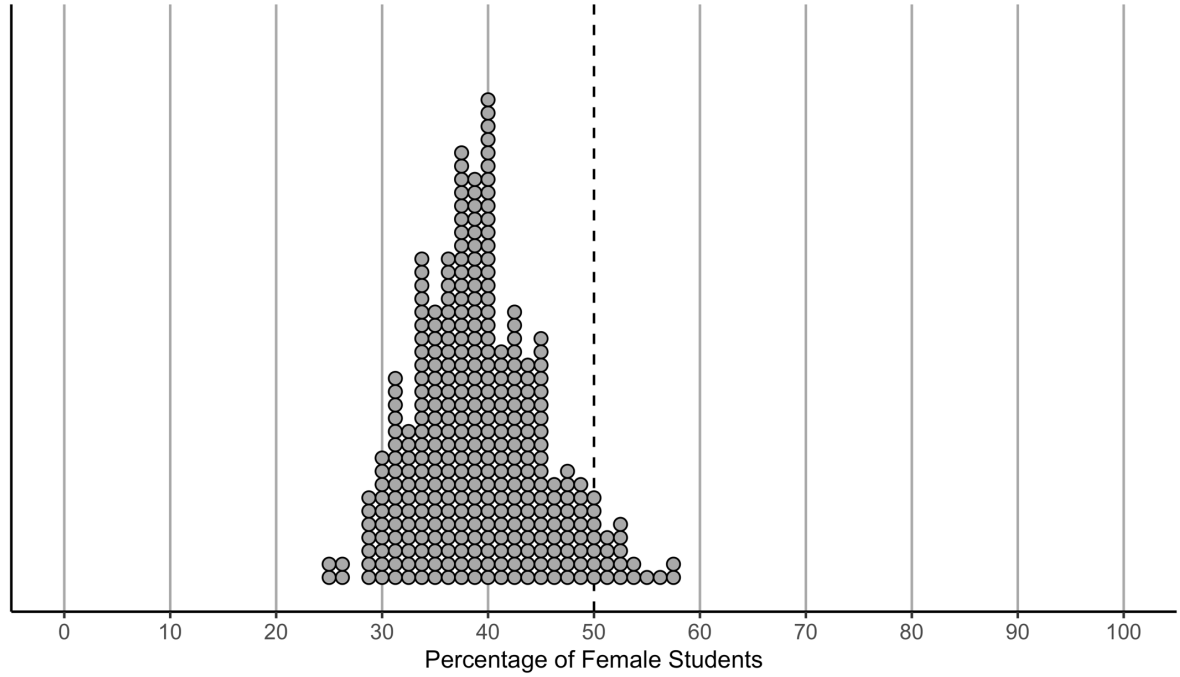
Note. Negative values indicate an advantage of female students, positive values an advantage of male students. See Tables S1 and S2 for more details on all included studies.

Figure 3. *PRISMA Flow Diagram*



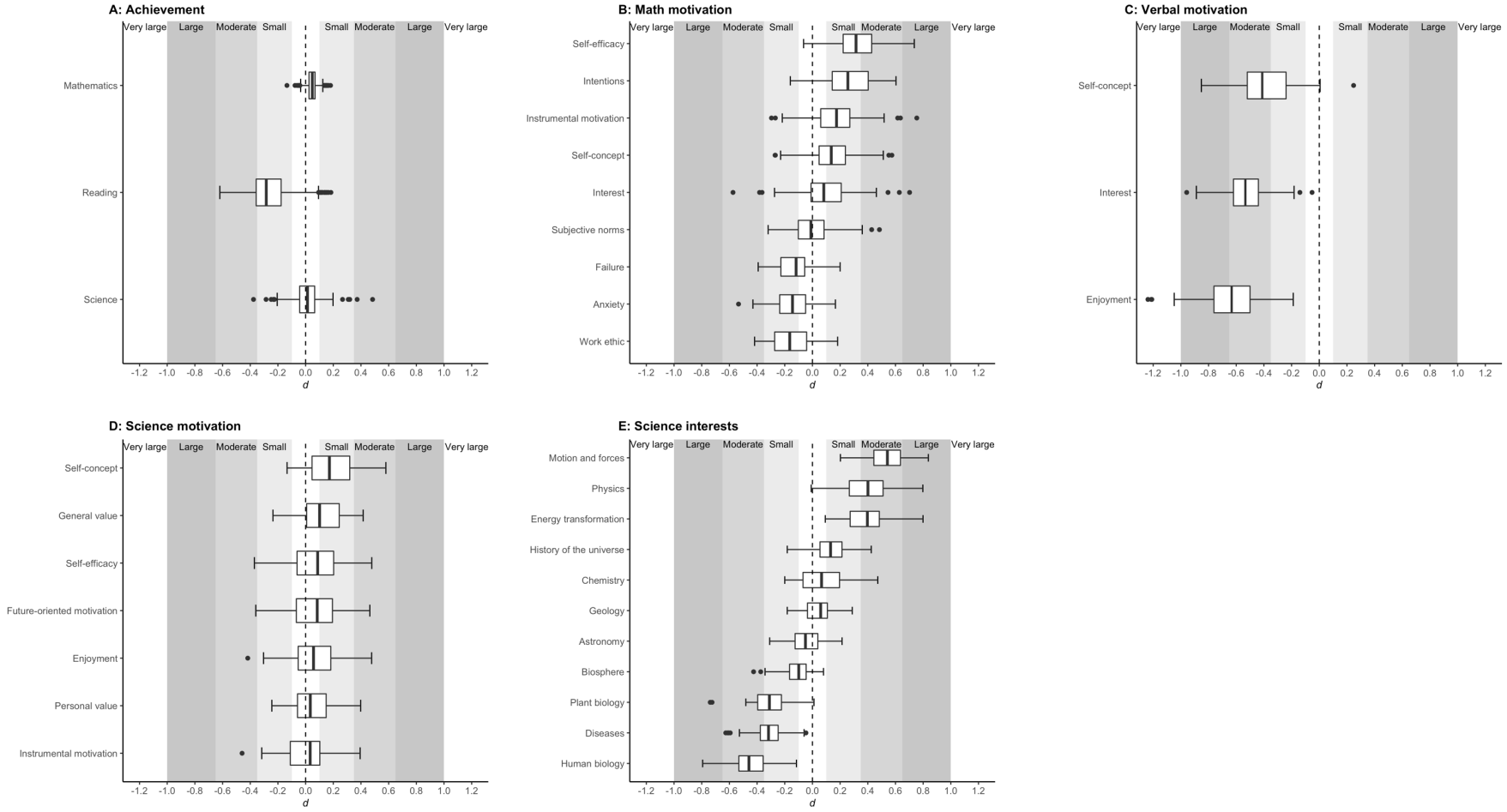
Note. Superscripts indicate which inclusion criterion led to the exclusion of the respective international large-scale assessment.

Figure 4. *Distribution of the Average Percentages of Female Students Belonging to the Top 5% in Mathematics Across 82 Countries (k =343)*



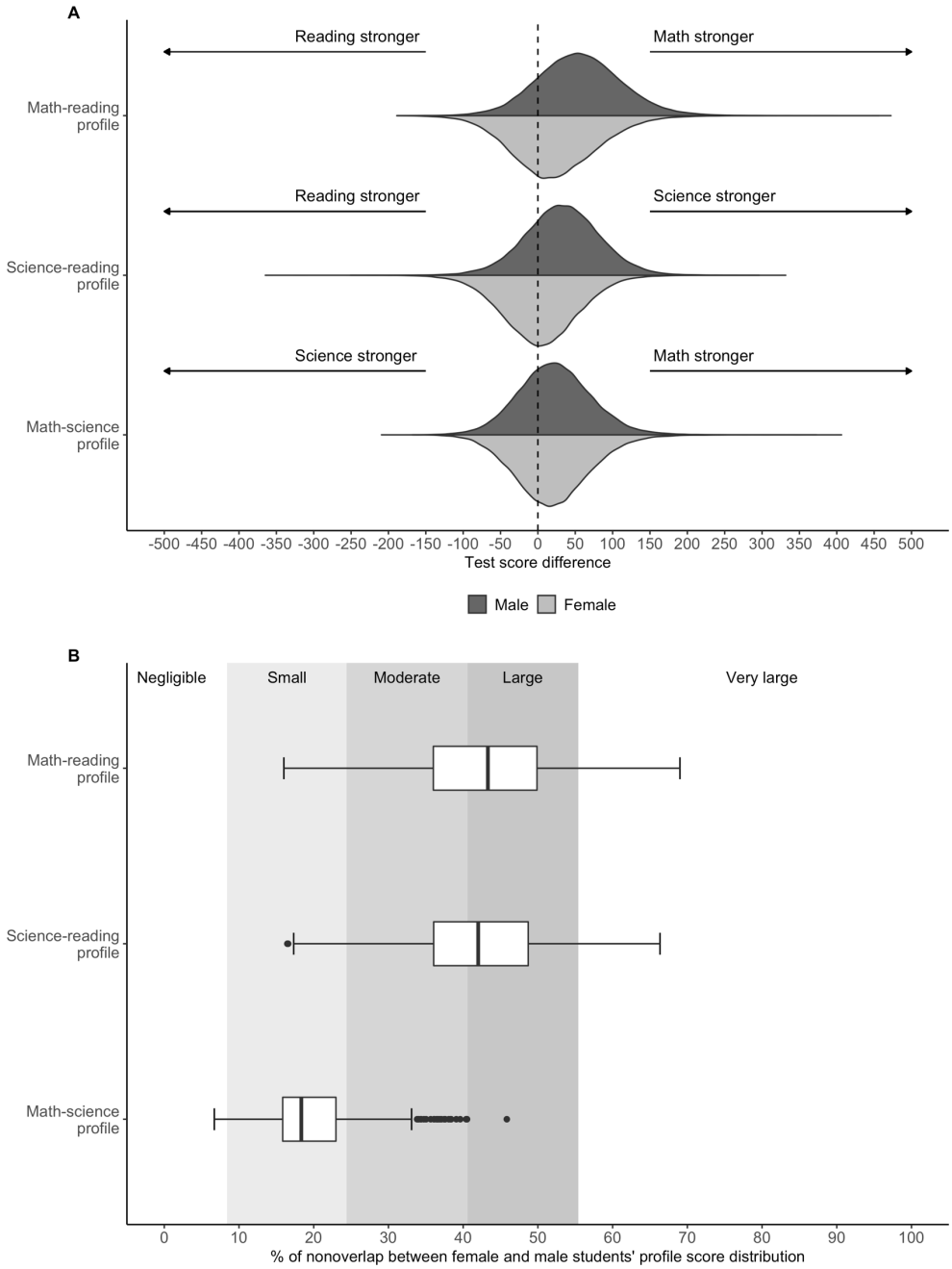
Note. Each dot represents one effect size. The vertical dashed line indicates gender parity.

Figure 5. *Distributions of Gender Differences (Cohen's d) in Top-Performing Math Students*



Note. Panel A: Achievement. Panel B: Math motivation. Panel C: Verbal motivation. Panel D: Science motivation. Panel E: Science interests. Boxplots comprise the median value of *d* (solid line in box), the 25th percentile (line below box), and the 75th percentile (line above box) of the *d* distribution. Negative values indicate an advantage of female students, positive values an advantage of male students.

Figure 6. *Distributions of Gender Differences in Achievement Profiles*



Note. Panel A: Distributions of achievement profile scores by gender in top-performing math students. $N = 115{,}481$. The figure is based on individual student data and the respective first plausible value. Panel B: Percentages of nonoverlap between female and male students' distributions of profile scores.

3

Study II

Nonlinear Relations Between Achievement and Academic Self-Concepts in Elementary and Secondary School: An Integrative Data Analysis Across 13 Countries

Keller, L., Preckel, F., & Brunner, M. (2020). Nonlinear Relations Between Achievement and Academic Self-Concepts in Elementary and Secondary School: An Integrative Data Analysis Across 13 Countries. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000533. The manuscript has been posted as a preprint on PsyArXiv (https://psyarxiv.com/8z563/).

3 Study II

Nonlinear Relations Between Achievement and Academic Self-Concepts in Elementary and Secondary School: An Integrative Data Analysis Across 13 Countries

Abstract

It is well-documented that academic achievement is associated with students' self-perceptions of their academic abilities, that is, their academic self-concepts. However, low-achieving students may apply self-protective strategies to maintain a favorable academic self-concept when evaluating their academic abilities. Consequently, the relation between achievement and academic self-concept might not be linear across the entire achievement continuum. Capitalizing on representative data from three large-scale assessments (i.e., TIMSS, PIRLS, PISA; $N = 470{,}804$), we conducted an integrative data analysis to address nonlinear trends in the relations between achievement and the corresponding self-concepts in mathematics and the verbal domain across 13 countries and two age groups (i.e., elementary and secondary school students). Polynomial and interrupted regression analyses showed nonlinear relations in secondary school students, demonstrating that the relations between achievement and the corresponding self-concepts were weaker for lower achieving students than for higher achieving students. Nonlinear effects were also present in younger students, but the pattern of results was rather heterogeneous. We discuss implications for theory as well as for the assessment and interpretation of self-concept.

Keywords: academic achievement, academic self-concept, mathematics, reading, nonlinear relations

Educational Impact and Implications Statement

The present study significantly advances the understanding of how performance on a standardized achievement test in a certain academic domain is related to students' corresponding academic self-concept. In representative student samples, we show that the relations between achievement and self-concepts in mathematics and the verbal domain can be better approximated by nonlinear relations, demonstrating weaker relations for lower achieving students than for higher achieving students in secondary school (and to some extent also in elementary school). Practitioners should be aware that there is no general linear trend between students' achievement and their corresponding academic self-concepts and should take this into consideration when assessing and interpreting students' academic self-concepts in counseling contexts.

Nonlinear Relations Between Achievement and Academic Self-Concepts in Elementary and Secondary School: An Integrative Data Analysis Across 13 Countries

Students' academic achievement is a major determinant of their academic self-concepts[6] (e.g., Harter, 2012; Marsh, 1986; Marsh & Craven, 2006; Möller et al., 2009; Shavelson et al., 1976; Trautwein & Möller, 2016). Typically, researchers implicitly assume linear relations between achievement and self-concepts (e.g., Huang, 2011; Marsh, 1986; Marsh & Hau, 2004; Möller et al., 2009, 2014; Skaalvik & Rankin, 1992). Linear relations between achievement and self-concepts imply that some constant amount of increase in achievement (e.g., an increase of one standard deviation) is associated with a constant increase in the corresponding self-concept (e.g., an increase of .30 standard deviations) across the entire achievement continuum. However, negative performance feedback can constitute a major threat to the self, which in turn may motivate lower achieving students to engage in self-protective strategies (Alicke & Sedikides, 2009). These self-protective strategies may weaken the potentially damaging effects of negative feedback on lower achieving students' self-concepts (Leary, 2007). Thus, it is likely that lower achieving students have more strongly inflated self-concepts than higher achieving students do. Consequently, linear models might not fully capture the relation between achievement and the corresponding self-concept for a large part of the student body. Rather, the relations between achievement and self-concepts may be better approximated by a nonlinear function that assumes weaker relations for lower achieving students and stronger positive relations for higher achieving students. Although nonlinear relations between achievement and self-concepts seem highly plausible, such relations have rarely been studied empirically. Therefore, the major aim of the present cross-national integrative data analysis was to substantially expand the body of knowledge on the

[6]Unless otherwise indicated, the term "achievement" will be used to indicate "academic achievement" and the term "self-concept" to indicate "academic self-concept" for the remainder of the article.

forms of the functional relations between achievement and self-concepts. As a major strength of this study, we tested the generalizability of our hypothesis across domains (i.e., mathematical and verbal), age groups (i.e., elementary and secondary school students), and countries. In addition, we replicated our results in different data sets. To this end, we drew on representative student samples from several cycles of international large-scale assessments: the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA). In doing so, the present results allowed us to draw reliable conclusions concerning theory (e.g., for models of self-concept formation) and the assessment and interpretation of self-concepts.

Integrative Data Analysis: A Tool for Enhancing the Robustness and Generalizability of Results

Scientific research is currently facing critical concerns about the replicability of results in many areas of psychology, education, and other fields (e.g., Ioannidis, 2005; Open Science Collaboration, 2015). Because any single study's results are influenced by that study's design, sample, measurement, and quantification of key constructs, we conducted a coordinated analysis to examine consistency in the findings across multiple data sets (Hofer & Piccinin, 2009). In a coordinated analysis—a form of integrative data analysis—several data sets that differ in samples and measures but assess similar constructs are first analyzed by applying an identical analysis protocol. Then, effect sizes are summarized across data sets using meta-analytic techniques to answer a specific research question (see also Graham et al., 2017). Similar to other tools that are used to synthesize research (e.g., meta-analysis or systematic literature reviews), coordinated analysis meets the need for a cumulative approach to scientific inquiry (Curran, 2009; Hunter & Schmidt, 1996; Meehl, 1978). Finally, the particular advantage of a coordinated analysis is that it can be used to obtain multiple

replications from independent data sets to thereby strengthen the confidence, robustness, and generalizability of findings.

Specifically, for the current investigation, we identified eight cycles from three international large-scale assessments that assessed both student achievement and students' self-concepts in mathematics and the verbal domain (i.e., TIMSS, PIRLS, and PISA) to examine the generalizability of nonlinear relations between achievement and the corresponding self-concepts across countries and time. In doing so, we took advantage of having access to individual participant data from representative probability samples of students. This allowed us to examine relations between achievement and self-concepts across the entire range of students' achievement and self-concepts because the data were not affected by range restrictions or selection bias. The present study involves the first integrative data analysis to use representative student data in the areas of student achievement and achievement motivation.

From Tools to Theories: Linear and Nonlinear Relations between Achievement and Self-Concepts

Theories are the starting point of scientific discoveries. Theory informs and guides scientific practices, such as the choice of methods that are applied to investigate a research question. However, theory is also shaped by scientific methods (tools-to-theories heuristic; Gigerenzer, 1991). This is exemplified in an allegory from Eddington (1939; see Cacioppo & Bernston, 1994) in which a hypothetical scientist attempted to determine the size of fish in the sea by sampling catches from a 2-inch net. After extensive fishing, the scientist did not find any fish smaller than 2 inches and therefore concluded that there were no fish smaller than 2 inches in the sea.

How is this allegory aligned with research on the relation between achievement and self-concept? Conducting a literature search in the data bases PsycINFO and ERIC for the

term “academic achievement” in combination with “academic self concept” resulted in a total of 1,229 peer-reviewed academic articles in PsycINFO and 1,005 peer-reviewed academic articles in ERIC published between 2009 and 2019, thus illustrating the large amount of research interest in this topic. Typically, researchers implicitly assume linear relations between achievement and self-concepts across the entire achievement continuum (e.g., Huang, 2011; Marsh, 1986; Marsh & Hau, 2004; Möller et al., 2009, 2014; Skaalvik & Rankin, 1992). In other words, most researchers have applied the “linear net” to study the relations between achievement and self-concepts. For example, a review of all studies (N = 66[7]) that were included in the two most recent meta-analyses on the relations between achievement and self-concept by Huang (2011) and Möller et al. (2009) revealed that all of these studies analyzed only linear relations between achievement and self-concept. Findings from the vast majority of these studies indicated positive relations between domain-specific achievement tests and corresponding self-concepts. The meta-analysis by Möller et al. (2009) reported a mean path coefficient for the relations between achievement on mathematics tests and mathematics self-concept of r = .57, and for the relations between achievement on verbal tests and verbal self-concept of r = .47. In the meta-analysis by Huang (2011), the longitudinal correlation between prior achievement (as measured with standardized tests) and subsequent self-concepts ranged from r = .19 to .23.

Only a few published studies have used the “nonlinear net” when examining the relation between achievement and self-concept. These studies can be grouped by the analytic strategy they applied. The first group of studies examined mean-level differences in self-concepts in relation to achievement differences by splitting the achievement continuum into discrete groups (i.e., low-, average-, and high-achieving students) using planned contrasts,

[7]The coding scheme and the results of the review can be accessed via the Open Science Framework (Soderberg, 2018) at https://osf.io/9cgzm/. Of note, one dissertation included in the meta-analysis by Huang (2011) could not be found and was therefore not included in our review.

ANOVAs, or multigroup comparisons in structural equation models (Möller & Pohlmann, 2010; Prast et al., 2018; Schurtz et al., 2014; Van der Beek et al., 2017). The second group of studies investigated the relation between achievement and self-concept by using polynomial regression analyses (Marsh, 2004; Marsh & Rowe, 1996).

A series of studies by Möller and Pohlmann (2010) and Schurtz et al. (2014) examined the relation between achievement and self-concept in the verbal domain (Möller & Pohlmann, 2010; Schurtz et al., 2014) and in mathematics (Schurtz et al., 2014) by comparing groups of low-, average-, and high-achieving students. Their results provided initial evidence that the strength of the relation between achievement and self-concept varies with students' level of achievement with stronger relations for higher achieving students and lower relations for lower achieving students in German (native language education; Möller & Pohlmann, 2010; Schurtz et al., 2014). However, the findings were heterogeneous for English as a foreign language and for mathematics (Schurtz et al., 2014).

It is important to note that estimates of the strength of a relation between two variables are biased when a continuous variable, such as student achievement, is split into discrete groups (MacCallum et al., 2002; Preacher et al., 2005). This limitation was overcome in the second group of studies that (a) tackled the achievement continuum as a whole and (b) integrated polynomial components into their analyses of the relation between the two constructs (Marsh, 2004; Marsh & Rowe, 1996). In a study that drew on a representative U.S. sample of male students in Grade 10, Marsh and Rowe (1996) found a positive quadratic relation ($\beta = 0.274$) between students' ability (as measured with an ability test) and their general academic self-concept. This indicated that increments in achievement were more strongly related to the self-concept of students with higher ability than to the self-concept of students with lower ability (Marsh & Rowe, 1996). In a representative Australian sample of 15-year-old students, Marsh (2004) found a positive quadratic relation ($\beta = 0.061$) between

general academic achievement (as measured with an achievement test) and general academic self-concept. This finding again implied that the relation between achievement and self-concept was stronger for higher achieving students and weaker for lower achieving students (Marsh, 2004). In the following section, we will present a possible mechanism that may underlie nonlinear relations between achievement and the corresponding self-concepts.

Self-Protection

One major reason for why the relations between achievement and self-concepts might vary as a function of individual student achievement is that being asked to evaluate one's own abilities in self-concept questionnaires may trigger self-protective strategies. When students are asked to evaluate their abilities with items such as "I learn quickly in mathematics" or "Reading is easy for me," they infer their abilities from achievement indicators, such as grades, tests, or other achievement-related feedback (Duckworth & Yeager, 2015; Marsh & Craven, 2006; Möller et al., 2009, 2014), and from social comparisons with their peers (Festinger, 1954; Huguet et al., 2009; Suls et al., 2002; Wheeler & Suls, 2005). Intuitively, one can expect that students with lower achievement in a specific domain should have lower evaluations of their abilities in this domain. However, a negative self-evaluation constitutes a major threat to the self. To protect their self-worth, low-achieving students are likely to engage in self-protective strategies that lead to more positive self-evaluations (Alicke & Sedikides, 2009). Self-protection originates from the assumption that people want to feel good or want to avoid feeling bad about themselves (Alicke & Sedikides, 2011). Self-protective strategies are aimed at avoiding, minimizing, and compensating for negative self-views (Hepper et al., 2010). For example, by changing the comparison group (i.e., comparing oneself with groups that have worse performances in the relevant dimensions; Tajfel & Turner, 1979), having poorer recall of self-threatening feedback than non-self-threatening feedback (Green & Sedikides, 2004), or downplaying the importance of a negative event

(Alicke & Sedikides, 2009), students can weaken the negative effect of their low achievement on their corresponding self-concept.

Empirical research has supported these theoretical propositions: In a study by Hacker, Bol, Horgan, and Rakow (2000), university students repeatedly took exams in a course and rated how well they thought they would perform before and after each test. Better performing students were able to increase the accuracy of their predictions as the semester proceeded, whereas poorly performing students showed no increase in their predictions despite the feedback the students received on their exam results (Hacker et al., 2000). Another study compared university students' self-reported grades from the previous semester with their actual grades and found that students with poorer grades overestimated their grades more than students with better grades did (Gramzow et al., 2003).

Furthermore, there is evidence that from the ages of 8 to 10, children already use self-protective strategies and that the need to use these strategies increases with students' (cognitive) development and life experience: In middle to late childhood, children develop many cognitive skills that enable them to integrate positive and negative information into their self-views (Harter, 2012). At about this age, most children have gathered experience with comparative grading practices and consequently with absolute failure or comparative performance feedback (Stipek & Daniels, 1988). However, achievement-related feedback becomes stricter and more varied in adolescence (Eccles et al., 1984). In addition, children's ability to draw social comparisons improves in middle to late childhood (Frey & Ruble, 1990; Harter, 2012; Ruble & Frey, 1990). Moreover, research has suggested that students at higher levels of cognitive maturation should show greater variability in their use of self-protective strategies (e.g., Alicke & Sedikides, 2011; Harter, 2012). In contrast to children, adolescents possess the cognitive skills that enable them to engage in attributional biases, such as attributing their successes to internal, stable characteristics (e.g., intelligence) and their

failures to external factors (e.g., the difficulty of a test; Harter, 2012). Furthermore, adolescents are able to protect their positive self-views by viewing their positive attributes as central and important and their negative characteristics as unimportant to their selves (Harter, 2012).

The Present Study

The overarching goal of the present study was to examine whether relations between achievement and corresponding self-concepts are nonlinear and to what extent the nonlinearity is generalizable across different domains, age groups, and countries. On the basis of theoretical considerations regarding strategies that support a positive self-view (e.g., Alicke & Sedikides, 2009), we expected that the relations between achievement and the corresponding self-concepts would be weaker for lower achieving students and stronger and positive for higher achieving students in the mathematical and verbal domains.

There are a few studies that have empirically supported this prediction to some extent (Marsh, 2004; Marsh & Rowe, 1996; Möller & Pohlmann, 2010; Schurtz et al., 2014). However, these studies have embodied several limitations. First, some of the studies used an analytical approach that can result in biased estimates (i.e., splitting the achievement continuum into discrete groups; Möller & Pohlmann, 2010; Schurtz et al., 2014). We capitalized on two analytical approaches that did not entail this limitation (i.e., quadratic and interrupted regressions). Second, other previous studies examined only general academic achievement and general academic self-concept (Marsh, 2004; Marsh & Rowe, 1996). However, self-concepts are highly domain-specific (e.g., Gogol et al., 2017). Therefore, we examined relations between achievement and self-concepts in mathematics and the verbal domain. Third, previous research has only investigated the relation between achievement and self-concept in secondary school students (Marsh, 2004; Marsh & Rowe, 1996). Yet, self-evaluations and the use of self-protective strategies are subject to age-related changes (e.g.,

Eccles et al.,1984; Guay et al., 2003; Harter, 2012; Marsh, 1989, 1990a; Wigfield & Eccles, 2002). Consequently, we examined the functional forms of the relations between achievement and self-concepts in different age groups (i.e., elementary and secondary school students). Fourth, a lesson learned from cross-cultural research is that "universality can never be assumed in advance" (Segall & Lonner, 1988, p. 1103; see also Henrich et al., 2010). It is not clear to what extent the findings from Australia (Marsh & Rowe, 1996) or the United States (Marsh, 2004) can be transferred to other countries. Thus, in the present study, we set out to examine the relations between achievement and self-concepts across 13 different countries. Fifth, given the current concerns about replicability (e.g., Ioannidis, 2005), we conducted an integrative data analysis (Curran & Hussong, 2009; Hofer & Piccinin, 2009) in which we applied the same analysis protocol to representative high-quality individual student data from three major large-scale assessments. This approach offered the advantage that we did not rely on findings based on a single study but rather integrated results across studies (Open Science Collaboration, 2015). To sum up, the results of the present study will significantly expand the body of knowledge on the functional forms of the relations between achievement and self-concepts in (a) different domains (i.e., mathematical and verbal), (b) age groups (i.e., elementary and secondary school students), and (c) countries.

Method

Samples

We used individual student data from three international large-scale assessments. The TIMSS, PIRLS, and PISA studies are designed and conducted to compare education systems worldwide and to provide policy makers, educators, researchers, and practitioners with reliable information about trends in mathematics, science, and/or reading achievement and learning contexts over time (Mullis, Martin, Kennedy, et al., 2009; Mullis, Martin, Ruddock, et al., 2009; Organisation for Economic Co-Operation and Development [OECD], 2013). In

TIMSS, students' skills and knowledge in mathematics and science are assessed every 4th year in Grades 4 and 8 (Mullis, Martin, Ruddock, et al., 2009). The PIRLS studies assess students' reading skills in Grade 4 every 5th year (Mullis, Martin, Kennedy, et al., 2009), and PISA assesses students' skills and knowledge in the core domains mathematics, reading, and science at the age of 15 every 3rd year (OECD, 2013). In every TIMSS, PIRLS, and PISA cycle, the school staff, students, and parents were informed about the nature of the test and the test date, and parental permission was secured if requested by the school or education system (Martin & Mullis, 2012; OECD, 2002).

Besides assessing students' skills and knowledge, several TIMSS, PIRLS, and PISA cycles measured students' domain-specific self-concepts. In stark contrast to other cycles from these large-scale assessments, the data on achievement and self-concepts in both the mathematical and verbal domains were collected from the same students in TIMSS/PIRLS 2011 (fourth-graders) and in PISA 2000 (15-year-olds). In other words, each student provided both mathematical and verbal data. Using the same samples of students when analyzing the domain-specificity of the relation between achievement and the corresponding self-concept measure ensured that the respective relations could only be influenced by the domain (mathematical or verbal) because other potentially confounding, person-related factors (e.g., socioeconomic background, cognitive ability, cohort membership) were controlled for. Given this strength of the data, we selected the countries that participated in the TIMSS/PIRLS 2011 cycle and the PISA 2000 cycle for our analyses (see Table 1). On the basis of this selection criterion, Romania would also have been included in the present analyses. However, preliminary analyses indicated severe problems with the PISA 2000 data such that there was a highly implausible relation between mathematics achievement and mathematics self-concept in Romania of $r = .00$. We therefore excluded the data from Romania from the present analyses.

To replicate our results, we additionally included the following assessment cycles in which both achievement and self-concept were measured in either mathematics or the verbal domain. We added cycle 2015 for TIMSS, cycle 2016 for PIRLS, and cycles 2003 and 2012 for PISA (covering the mathematics domain) to our analyses. In PISA 2000, a mathematics achievement score was provided for (a random subsample of) 56% of the students in the public use file. This resulted in a smaller number of students compared with the other PISA cycles where a mathematics achievement score was provided for all students. Only eight (out of 13) countries participated in the TIMSS assessment for Grade 8 in 2011 and 2015 (see Table 1). Austria did not participate in the TIMSS 2015 Grade 4 assessment. Because PISA assessed students' self-concept in the verbal domain in the year 2000 cycle only, it was not possible to replicate these results. Finally, some information on students' gender was missing. In sum, 328 students were excluded from the analyses because information on their gender was missing (most of these cases came from PISA 2000 from which 239 students were excluded).

TIMSS, PIRLS, and PISA capitalize on a two-stage stratified sampling design to achieve representative probability samples (a detailed description of the sampling procedures can be found in the supplemental online materials [SOM], which can be accessed via the Open Science Framework (Soderberg, 2018) at https://osf.io/9cgzm/). All in all, our analyses were based on representative student samples comprising data from a total of 470,804 students in 23,307 classes or schools in which 50% of the students were female, and the mean age ranged from 10.30 years (TIMSS/PIRLS 2011, Grade 4) to 15.78 years (PISA 2003; see Table 1). Sample sizes, percentages of female students, and numbers of participating schools and classes for each assessment cycle and each country can be found in Tables S1 and S2 in the SOM. Students' mean age in every assessment cycle is shown in Tables S3 to S10 in the SOM.

Measures

Achievement. Student achievement in mathematics and reading was measured with the standardized tests used in the respective TIMSS, PIRLS, and PISA assessments. In Grade 4, the TIMSS mathematics assessment contained the content domains number, geometric shapes and measure, and data display; in Grade 8, the assessment covered the domains number, algebra, geometry, and data and chance (Mullis, Martin, Ruddock, et al., 2009; Mullis & Martin, 2013). The PIRLS assessment framework focused on the two overarching purposes for reading in Grade 4: literary experience, and acquiring and using information (Mullis & Martin, 2015; Mullis, Martin, Kennedy, et al., 2009). In PISA cycles 2003 and 2012, mathematical content knowledge was assessed in four categories: change and relationships, space and shape, quantity, and uncertainty and data (OECD, 2003, 2013). In PISA 2000, the assessment covered just two categories: space and shape, and change and relationships. Reading literacy was assessed in three different categories in PISA: the abilities to access and retrieve information, integrate information and interpret texts, and reflect upon and evaluate texts (OECD, 2000).

The achievement assessments in TIMSS, PIRLS, and PISA were designed to pursue different goals: Whereas the TIMSS and PIRLS assessments are classroom- and curriculum-based, PISA focuses on literacy concepts, that is, students' abilities to apply their skills and knowledge in mathematics and reading to everyday life problems. Thus, at least for mathematics, for which secondary school students are tested in both PISA and TIMSS (Grade 8), PISA is considered a more challenging assessment (e.g., Else-Quest et al., 2010).

In all TIMSS, PIRLS, and PISA assessments, the achievement scores for mathematics and reading were scaled to have an international mean of 500 points and a standard deviation of 100 points (Martin et al., 2016, 2017; OECD, 2014). All achievement scales underwent extensive field testing before being implemented in the respective assessments (Martin et al.,

2016, 2017; OECD, 2014). Means and standard deviations for the achievement measures from all countries and assessments are reported in Tables S3 to S10 in the SOM. Table 2 shows the median reliability of mathematics and reading achievement and self-concept scales in each assessment. The reliabilities were satisfactory in all assessment cycles in all participating countries. The reliabilities of the achievement and self-concept scales in each country can be found in Tables S11 to S14 in the SOM.

To estimate students' achievement scores, TIMSS, PIRLS, and PISA used plausible values. Plausible values are representations of the range of abilities a student may reasonably have. Hence, TIMSS, PIRLS, and PISA estimated probability distributions for each student's true achievement score. Applying plausible values offers the methodological advantage of unbiased estimates of population parameters (e.g., means and standard deviations). Overall, TIMSS, PIRLS, and PISA provided five plausible values for each achievement scale.

Academic self-concept. To assess students' self-concepts in mathematics (TIMSS, PISA) and reading (PIRLS) or the verbal domain (PISA), students used 4-point rating scales to provide answers in all assessments with higher scores indicating a higher self-concept in the respective domain. The mathematics and verbal self-concept items used in the TIMSS, PIRLS, and PISA assessments rely on established items whose wordings were identical or similar to well-established and researched questionnaires such as the Self-Description Questionnaire (SDQ; e.g., Marsh, 1990b; Byrne, 1996, 2002; Marsh et al., 2006): e.g., "I learn quickly in [domain]" (used in the SDQ II), "I get good marks in [domain]," "I learn [domain] quickly" (used in PISA 2000, 2003, 2012), or "I usually do well in [domain]" (used in PIRLS 2011, 2016; TIMSS 2011, 2015). All self-concept scales underwent extensive field testing before being implemented in the respective assessments (Martin et al., 2016, 2017; OECD, 2014). To create scale scores, the ratings for the single self-concept items were averaged. The reliabilities for all the TIMSS, PIRLS, and PISA self-concept scales were

consistently satisfactory in all cycles and across all participating countries (see Tables S11 to S14). Means, standard deviations (Tables S3 to S10), item wording, and response scales (Tables S15 to S18) can be found in the SOM.

Data Analysis

As to be expected in any (large-scale) study, there were some missing data. Specifically, 4% to 8% of the self-concept data were missing in each sample, with one exception of 37% missing from PISA 2012 because students' responses were missing by design in this cycle (OECD, 2014). The country-specific percentage of missing information in each assessment and cycle can be found in Tables S3 to S10 in the SOM. Missing data in students' self-concepts were handled by applying nested multiple imputation (e.g., Weirich et al., 2014) using the R package "miceadds" (version 2.14-26; Robitzsch et al., 2018). For each plausible value representing students' achievement, we imputed the missing self-concept data five times, yielding 25 nested imputations for every data set. We imputed the missing data in students' self-concepts separately for female and male students because research has shown that female students typically report higher self-concepts in the verbal domain but lower self-concepts in mathematics in comparison with their male counterparts (e.g., Jacobs et al., 2002; Marsh & Yeung, 1998; Skaalvik, & Skaalvik, 2004). All analyses on the relation between achievement and self-concept in this study were computed 25 times, and then the results were integrated using standard procedures to obtain an average estimate as well as corresponding standard errors (Martin et al., 2016, 2017; OECD, 2014).

We conducted an integrative data analysis (see Curran & Hussong, 2009; Hofer & Piccinin, 2009) where we applied the same analysis protocol to investigate the functional forms of the relations between achievement and self-concept in mathematics and the verbal domain. Specifically, we proceeded in two ways. First, we ran domain-specific regression models for every country in every assessment and every cycle with achievement in

mathematics or reading as the predictor variable and mathematics or reading/verbal self-concept as the outcome variable. The linear model specified a linear relation between achievement and self-concept; the quadratic model specified a nonlinear relation and included a linear term and a quadratic term. To analyze the regression models, we used the statistical software R (version 3.5.1; R Core Team, 2018) and the R package "BIFIEsurvey" (version 2.18-6; BIFIE, 2018). Before computing the regression models, we standardized (M = 0.00, SD = 1.00) students' achievement and self-concept scores around each country mean in every TIMSS, PIRLS, and PISA cycle, respectively. In doing so, we (a) facilitated the comparison of regression parameters between achievement and self-concept scales across assessments and (b) removed nonessential multicollinearity, that is, the multicollinearity between a variable and the higher order function of the same variable (here: achievement) that exists merely because of the scaling (nonzero mean) of the variable (Cohen et al., 2003). To evaluate the significance of the linear and quadratic regression coefficients, we computed 95% confidence intervals (CIs). If the 95% CI did not contain 0, the changes in achievement were interpreted as significantly related to changes in the corresponding self-concept (Cohen et al., 2003). The regression coefficients for single countries and cycles can be found in Tables S19 to S23 in the SOM.

Second, we ran interrupted regressions using the two-lines test (Simonsohn, 2018) with the R packages "mgcv" (version 1.8-24; Wood, 2017) and "survey" (version 3.34; Lumley, 2004) for every assessment and cycle. We used the interrupted regressions to explore other possible nonlinear trends apart from quadratic effects between students' domain-specific achievement and the corresponding self-concept. In the two-lines test, an interrupted regression for low and high achievement scores is estimated by dividing the achievement continuum into two segments (Figure 1a). The breakpoint between segments is determined by the so-called Robin Hood algorithm, which maximizes the precision with which the

regression parameters β_1 and β_2 are estimated. β_1 and β_2 describe the linear relation between achievement and self-concept within each segment, respectively. To this end, the algorithm allocates the observations between the two segments such that the standard errors of the regression parameters β_1 and β_2 are minimized without changing the values of β_1 and β_2 too much (Simonsohn, 2018). If β_1 and β_2 are equal, this would argue for a linear relation between achievement and the corresponding self-concept. Consequently, if β_1 and β_2 are not equal (e.g., β_1 is notably smaller than β_2), this would argue for a nonlinear relation between the two constructs. The breakpoint is particularly important for the latter case. Because the distribution of students' achievement can be well-approximated by a (standard) normal distribution within each country (Figure 1a), we can estimate the proportion of students for which a single linear regression coefficient does not fully capture the relation between achievement and the corresponding self-concept. To provide a conservative estimate, we computed the proportion such that it represented the students whose achievement scores were located in the segment with fewer observations below or above the breakpoint, respectively. To this end, we computed the proportion of students under a standard normal distribution who had achievement scores below the breakpoint if the breakpoint was less than or equal to 0. For example, a breakpoint of z = -1.04 (as depicted in Figure 1a) suggests that for about 15% of the students, the relation between achievement and self-concept was not accurately represented by a single linear regression coefficient. If the breakpoint was greater than 0, we computed the proportion of students under a standard normal distribution who had achievement scores above the breakpoint.

For both analytic strategies, we used meta-analytic techniques to integrate the results across 13 countries and several cycles for each domain and student group (i.e., Grade 4, Grade 8, 15-year-olds), respectively. In accordance with the analysis strategy proposed by Cheung and Jak (2016) for big data, we proceeded in two steps. First, we analyzed the

individual student data according to the detailed TIMSS, PIRLS, and PISA guidelines for statistical analyses on how to apply population weights, compute standard errors, and compute regression coefficients with plausible values as achievement indicators (Martin & Mullis, 2012; OECD, 2014). Second, to obtain a single estimator within each student group (Grade 4, Grade 8, 15-year-olds), we computed weighted mean regression coefficients across countries (see Tables S19 to S23 and S29 to S33 in the SOM). To this end, we used three-level random-effects models with maximum likelihood estimation to account for the dependencies between regression coefficients when the regression coefficients were obtained for several cycles within countries (i.e., in TIMSS, PIRLS, PISA mathematics). Because verbal self-concepts were only measured in one PISA cycle (PISA 2000), there was no dependency in regression coefficients within countries. Thus, we conducted a two-level random-effects model with maximum likelihood estimation in this case. We applied random-effects models to allow the true effect to vary between (and within) countries (Borenstein et al., 2009; Cheung, 2015). For the three-level random-effects models, estimates of the variability in the regression coefficients defined Level 1. Level 2 captured variability in regression coefficients between assessment cycles within countries, and Level 3 captured variability in regression coefficients between countries. For the two-level random-effects model, Level 1 captured the estimates of the variability in the regression coefficients, and Level 2 captured the variability in regression coefficients between countries. We used the R package "metaSEM" (version 1.2.2; Cheung, 2015) to analyze the two- and three-level random-effects models.

We computed three statistics to assess the heterogeneity of the effect sizes: I^2, τ, and Q (Borenstein et al., 2009). Higgins and Thompson's (2002) measure of heterogeneity I^2 represents the proportion of observed heterogeneity that is real and not due to random noise. I^2 has a range of 0% to 100%, such that 30% to 60% may represent moderate heterogeneity,

50% to 90% may represent substantial heterogeneity, and 75% to 100% may represent considerable heterogeneity (Higgins & Green, 2011; Higgins & Thompson, 2002). In our study, I^2_{Level2} captured the heterogeneity between cycles within countries, I^2_{Level3} captured the heterogeneity between countries, and I^2_{total} captured the total heterogeneity (i.e., the sum of I^2_{Level2} and I^2_{Level3}). We also report the standard deviation of the regression coefficients τ (see Borenstein et al., 2009) to estimate heterogeneity in regression coefficients between cycles within countries (τ_{Level2}) and between countries (τ_{Level3}) plus the total heterogeneity (τ_{total}; $\tau_{\text{total}} = \sqrt{\tau^2_{\text{Level2}} + \tau^2_{\text{Level3}}}$). Finally, the Q test statistic (introduced by Cochran, 1954) is computed by summing the squared deviations of each individual effect size estimate from the corresponding overall (average) effect estimate where individual effect sizes are weighted by their sampling variance (Huedo-Medina et al., 2006). A statistically significant Q value indicates effect size heterogeneity (Borenstein et al., 2009).

All figures were produced using the R package "ggplot2" (version 3.1.0; Wickham, 2009). The R code for reproducing the results and figures from the present study can be accessed via the Open Science Framework (Soderberg, 2018) at https://osf.io/9cgzm/.

Results

Linear and Quadratic Regressions

Mathematics. Figure 2 displays the standardized linear and quadratic relations between mathematics achievement and mathematics self-concept. The results obtained for the linear model indicated that math achievement and self-concept were positively related for students in Grades 4 and 8 and for students at age 15: The higher students' achievement, the higher their corresponding self-concept.[8]

[8] In line with previous research (e.g., Marsh, 1986), in almost all countries, male students reported a higher self-concept in mathematics than female students did, whereas female students reported a higher self-concept in the verbal domain than male students did. The addition of students' gender as a predictor of their domain-specific self-concept did not change the functional forms of the relations between achievement and the corresponding

The results obtained for the quadratic model showed significant quadratic relations between achievement and self-concept in mathematics across countries. As shown in Figure 2, we observed positive quadratic relations in the group of 15-year-old students (mean β = 0.12, 95% CI [0.10, 0.14]) and Grade 8 students (mean β = 0.12, 95% CI [0.08, 0.15]), implying that the increase in students' mathematics self-concept was weaker for lower achieving students than for higher achieving students. Importantly, these quadratic relations were found across cycles and countries with very few exceptions (1 in 39 cases *ns* for PISA and 2 in 16 *ns* for TIMSS; Figures 2 and 3). The linear and quadratic terms demonstrated significant heterogeneity (Tables 3).

In younger students, the mean quadratic relation between mathematics achievement and mathematics self-concept was positive (mean β = 0.04, 95% CI [0.03, 0.06]). Similar to the group of older students, this indicated that the increase in students' mathematics self-concept in Grade 4 was lower for lower achieving students than it was for higher achieving students. However, across countries, positive quadratic relations were found less consistently for younger students than for older students; in 13 out of 25 cases, the quadratic relations were significant (see Table S19 in the SOM). Heterogeneity analyses indicated significant heterogeneity in the linear and quadratic relations for students in Grade 4 (Table 3).

Verbal domain. Figure 4 shows the standardized linear and quadratic relations between reading achievement and reading self-concept for students in Grade 4 and between reading achievement and verbal self-concept in 15-year-olds. The direction of the quadratic relations differed between younger and older students. Whereas the average quadratic relation between reading achievement and verbal self-concept was significantly positive in the group of 15-year-olds (mean β = 0.05, 95% CI [0.03, 0.07], Figure 4), it was significantly negative in Grade 4 (mean β = -0.02, 95% CI [-0.03, -0.01]). For 15-year-old students, reading

self-concepts. Results of the linear and quadratic regression models when controlling for students' gender are presented in Tables S24 to S28 in the SOM.

achievement was more strongly related to verbal self-concept for higher achieving students than for lower achieving students. The negative quadratic relation found in younger students implied that reading achievement and self-concept were to some extent more strongly related for lower achieving students than for higher achieving students.

Across countries, in the group of 15-year-olds, 6 out of 13 quadratic relations were significant. For fourth-graders, 5 out of 26 quadratic relations were significant (Figures 4 and 5, and Tables S22 and S23 in the SOM). For both fourth-graders and 15-year-olds, we observed significant heterogeneity in the linear and quadratic relations (Table 4).

Interrupted Regressions and Two-Lines Tests

Mathematics. The results from the interrupted regression models for 15-year-olds revealed that mathematics achievement was not significantly related to mathematics self-concept for lower achieving students, but the relation was significantly positive for higher achieving students (mean $\beta_1 = 0.03$, 95% CI [-0.01, 0.07]; mean $\beta_2 = 0.47$, 95% CI [0.41, 0.53]; Figure 1b). This finding held for the vast majority of countries across three assessment cycles (Figure 1b). We observed significant heterogeneity in both regression coefficients (i.e., β_1 and β_2; Table 5). The mean percentage of students for whom a common linear model did not accurately describe the relation between achievement and self-concept in mathematics was 16%.

For students in Grade 8, the relation between achievement and self-concept in mathematics was stronger for higher achieving students than for lower achieving students (mean $\beta_1 = 0.25$, 95% CI [0.08, 0.42]; mean $\beta_2 = 0.58$, 95% CI [0.50, 0.66]; Figure 1b). We observed significant heterogeneity in both regression coefficients (i.e., β_1 and β_2; Table 5). In about half of the countries, β_1 (as obtained for lower achieving students) and β_2 (as obtained for higher achieving students) were both positive; in the other half of the countries, β_1 was close to zero, but β_2 was positive. The mean percentage of students for whom a single linear

regression coefficient did not fully capture the relation between achievement and self-concept in mathematics was 12%.

For Grade 4, the achievement-related increase in mathematics self-concept was on average almost identical in magnitude for higher and lower achieving students (mean β_1 = 0.32, 95% CI [0.26, 0.39]; mean β_2 = 0.34, 95% CI [0.29, 0.38]; Figure 1b). This implied that the achievement-related increase in self-concept did not vary across the achievement distribution for students in Grade 4. Both regression coefficients exhibited significant heterogeneity that was entirely located between assessment cycles within countries (Table 5). There were three countries (i.e., Australia, Portugal, and Sweden in TIMSS 2011) in which the relation between mathematics achievement and mathematics self-concept was close to zero for lower achieving students but positive for higher achieving students. For an average of 17% of the students, a common linear model did not accurately describe the relation between achievement and self-concept in mathematics.

Verbal domain. The achievement-related increase in verbal self-concept in 15-year-olds was somewhat stronger for higher than for lower achieving students (mean β_1 = 0.20, 95% CI [0.14, 0.26]; mean β_2 = 0.32, 95% CI [0.25, 0.39]; Figure 1b). For some countries, the relation between reading achievement and verbal self-concept was close to zero for lower achieving students but positive for higher achieving students, whereas in other countries, the relation was positive for both lower and higher achieving students (Figure 1b). Heterogeneity measures showed that the magnitudes of the regression coefficients varied significantly (Table 6). The mean percentage of students for whom a single linear regression coefficient did not fully capture the relation between reading achievement and verbal self-concept was 22%.

The relation between achievement and self-concept in reading was stronger for lower achieving students than for higher achieving students in Grade 4 (mean β_1 = 0.45, 95% CI

[0.42, 0.48]; mean $\beta_2 = 0.27$, 95% CI [0.22, 0.31]; Figure 1b). Heterogeneity measures showed that the magnitudes of β_1 varied significantly, whereas the magnitudes of β_2 did not (Table 6). For an average of 13% of the students, a common linear model did not describe the relation between achievement and self-concept in reading very well. The results for single countries and cycles can be found in Tables S29 to S33 in the SOM.

Discussion

The major aim of the present integrative data analysis was to investigate whether the relations between achievement and self-concept are nonlinear and to what extent the nonlinearity can be generalized across (a) different domains (i.e., mathematics and verbal), (b) different age groups (i.e., elementary and secondary school students), and (c) 13 different countries. Most previous research applied a "linear net" to capture relations between achievement and self-concept measures. Yet, this approach fails to capture potential nonlinear relations between the two constructs and thus might not accurately describe the relations between achievement and self-concepts for a considerable proportion of the student body. In the present integrative data analysis, we capitalized on representative individual student data from eight assessment cycles of three major educational large-scale studies (i.e., TIMMS, PIRLS, PISA) and applied polynomial and interrupted regression analyses as "nonlinear nets." Our findings provided strong evidence of nonlinear relations between achievement and self-concepts for students in secondary schools in mathematical and verbal domains. Nonlinear effects were also present in younger students, but the result patterns were rather heterogeneous across countries and applied methods.

Implications for Theories of Self-Concept Formation

For secondary school students, the relations between achievement and the corresponding self-concepts in mathematics and the verbal domain were weaker for lower achieving students than for higher achieving students. This conclusion was supported by the quadratic and

interrupted regression analyses. For 15-year-olds in mathematics, the relation between mathematics achievement and mathematics self-concept was even close to zero for lower achieving students but positive for higher achieving students. Importantly, we replicated the positive quadratic relation between achievement and self-concept in mathematics in secondary school students over three (PISA) and two (TIMSS) assessment cycles, underpinning the robustness of our findings (achievement and self-concept in the verbal domain were only assessed once in PISA 2000, and therefore, it was not possible to replicate the effect). Similar nonlinear findings have been reported for the association between global achievement and global academic self-concept for students in Grade 10 and 15-year-olds (Marsh, 2004; Marsh & Rowe, 1996).

One plausible explanation for the finding that lower achieving students' self-concepts in mathematics and the verbal domain were only weakly related to their corresponding achievement is that lower achieving students are more likely to apply self-protective strategies to prevent the damaging effects of negative performance feedback on their self-views as has been shown, for example, in studies by Gramzow et al. (2003) and Hacker et al. (2000). Besides self-protective motives, students' self-enhancing motives might also contribute to the nonlinearity between achievement and the corresponding self-concepts to some extent. For instance, self-enhancing motives might motivate better performing students to increase the positivity of their self-concepts in response to positive performance feedback. However, there is evidence that lower achieving students are more inclined to apply self-serving strategies compared with higher achieving students (e.g., Baumeister et al., 2001; Gramzow et al., 2003; Hacker et al., 2000). This underscores the plausibility of self-protection as one driving mechanism that leads to nonlinear relations between achievement and self-concepts.

Nonlinear effects were also present in younger students, but the result patterns were rather heterogeneous. On average, we found a positive quadratic relation between

achievement and self-concept in mathematics. However, in more than half of the countries, the quadratic regression coefficient was not significant. Further, the interrupted regression analyses indicated that the relations between achievement and self-concept in mathematics were positive and almost identical in magnitude for lower and higher achieving students.

In the verbal domain, we found a negative quadratic relation, indicating that reading achievement and reading self-concept were less strongly related for higher achieving students in elementary school than for lower achieving students. The results from the interrupted regressions confirmed these findings. On the one hand, these findings could support the assumption that there are age differences in nonlinear relations. On the other hand, the nonlinear effects in the verbal domain could differ for younger and older students because younger students were asked to rate their reading self-concept in PIRLS, whereas older students rated their national language class self-concept in PISA. Typically, students do not receive specific grades for their reading performance, but instead they receive a global assessment of their performance in diverse areas of their national language class (e.g., writing, reading, grammar). Because grades are usually the most salient source of performance feedback for students, younger students have probably gained less comparative feedback experience that they can take into account when evaluating their reading performance. Consequently, we cannot distinguish whether the differences between younger and older students occurred for developmental reasons or due to differences in the assessed constructs (reading vs. verbal self-concept). More research and data in which verbal self-concepts are measured consistently across age groups are needed to answer this question.

In sum, the present results support two major conclusions. First, the results of the present study indicated that a linear regression model could not fully capture the relations between achievement and the corresponding self-concepts in mathematics and the verbal domain for a substantial proportion of the student body in secondary school. Considering the

country-specific breakpoints from the interrupted regression analyses, the "substantial proportion" ranged from, on average, 12% (TIMSS, Grade 8) to 22% (PISA, verbal domain) of the students (see Tables S29 to S33). Consequently, it would be advisable to refine current models of self-concept formation that assume that the relations between achievement and self-concepts are solely linear. For example, the internal/external frame of reference model (Marsh, 1986) states that students' self-concepts are formed by two processes: (a) an external (social) comparison process in which students compare their own achievement in a domain with their classmates' achievement in the same domain and (b) an internal comparison process in which students compare their own achievement in different domains. For example, to better capture the relation between achievement and self-concept in this model, a quadratic term could be included in the regression of self-concept on achievement in the same domain when modeling the external comparison process.

Second, the findings of the present cross-sectional study were mixed regarding age differences in nonlinear relations between achievement and self-concepts in mathematics and the verbal domain. Given the great heterogeneity in results as observed for elementary school students, we strongly recommend that nonlinear models also be specified for this student group. By using the "nonlinear net," researchers will avoid missing any nonlinear relations between achievement and self-concept. Doing so will also improve current theories on self-concept formation because this approach will eventually help to identify the boundary conditions and moderating factors that lead to linear relations for elementary school students in some countries and nonlinear relations in others.

Implications for the Assessment of Self-Concepts

In addition to research contexts, self-concepts are also assessed in guidance and counseling contexts. Our findings showed that lower achieving secondary school students tended to overestimate their academic abilities in mathematics and the verbal domain relative

to higher achieving students in these domains. The tendency to overestimate one's own abilities can introduce a systematic bias when lower achieving students seek guidance or counseling, and school counselors, school psychologists, or teachers make recommendations for their further educational or occupational pathways on the basis of students' scores on self-concept measures.

In the following, we suggest three methodological approaches that can be applied in practice to derive more nuanced interpretations of students' scores on self-concept scales. One approach could be to compare a student's self-reported self-concept with this student's actual performance (e.g., the student's grades or his or her achievement of performance benchmarks on tests). Another way could be to estimate students' tendency to engage in self-protection. This could be done, for example, by measuring how strongly the self-view of a (low-achieving) student diverges from others' (e.g., classmates') assessments of this student's academic abilities in a particular domain (e.g., Krueger & Wright, 2011). A self-protective tendency would be apparent when students evaluated their own abilities more positively than others did. Finally, students' self-protective tendency could also be measured by using self-report scales that directly assess students' application of self-protective strategies as suggested, for instance, by Hepper et al. (2010). These self-protection parameters could then be included when interpreting students' observed scores on self-concept scales.

Limitations and Outlook

Although this integrative data analysis significantly expands the body of knowledge on nonlinear relations between achievement and self-concepts in different domains, age groups, and countries, we see three limitations that should be addressed in future research.

Causal mechanisms. We drew on self-protection as the theoretical basis (e.g., Alicke & Sedikides, 2009) for expecting nonlinear relations between achievement and the corresponding self-concepts. However, given the limitations of our study design, we could not

conclude that self-protective strategies caused the nonlinear relations found in this study. Other research designs would be needed to do so. If self-protective strategies caused the nonlinear relation between achievement and self-concept, we would expect that experimentally enhancing students' motives to apply these strategies would lead to a stronger degree of nonlinearity between achievement and self-concept. For example, performance feedback is known to activate the motive to apply self-protective strategies (Alike & Sedikides, 2009). Hence, using experimental manipulations, students could be randomly assigned to groups that either receive feedback or do not receive feedback on their actual performance on a standardized achievement test before they evaluate their self-concepts. We would expect the degree of nonlinearity between achievement and self-concepts (as expressed in the quadratic or interrupted regression models) to increase for students in the feedback group because the experimental manipulation has direct consequences for how students evaluate their achievement (e.g., performance feedback [vs. no performance feedback] should intensify the threat to the self for lower achieving students). Another nonexperimental study design that could be applied to approximate the causal mechanism is to examine how (individual) differences in the application of self-protective strategies as measured with questionnaires (see Hepper et al., 2010) may moderate the functional form between achievement and self-concept.

Further influences on the functional form. By investigating students' self-concepts with data from TIMSS, PIRLS, and PISA, we were bound to the 4-point rating scales that were applied in these large-scale assessments. If a broader rating scale (e.g., a 6- or 7-point rating scale) had been used to measure the self-concepts, it may have been possible to find additional inflection points (e.g., at the high end of the achievement continuum). One possible reason for this is that the number of response categories seems to influence individuals' response behavior (Weijters et al., 2010). For example, Weijters et al. (2010) found that the

tendency to choose extreme anchor points on rating scales was reduced when more response categories were added to a rating scale. Future research should investigate the extent to which response format influences the functional form of relations between achievement and corresponding self-concepts. In addition, future studies should take into account possible response styles. For example, a study by Buckley (2009) provided initial evidence that adjusting for students' response biases (acquiescence, disacquiescence, extreme response styles, noncontingent responding) tended to result in stronger nonlinear relations between science achievement and science attitude scales in PISA 2006. Moreover, the anchoring vignette approach is a promising way to control for response style effects (Bolt et al., 2014; He et al., 2017) and was first applied in PISA 2012 (Kyllonen & Bertling, 2014; OECD, 2014). However, the anchoring vignette approach for self-concept scales still needs to be refined (e.g., He et al., 2017).

Heterogeneity in effects. Heterogeneity analyses revealed considerable heterogeneity in the linear and quadratic relations as well as in almost all interrupted regression coefficients in all domains and especially in older students. In particular, even the magnitudes of the linear relations were subject to variation. This may have (at least) two probable reasons. One explanation could be that relations between achievement and self-concept are stronger for some countries (and/or successive student cohorts within countries) than for others because the curriculum in mathematics or the verbal domain better matches the conceptual frameworks defined and measured in PISA, TIMSS, or PIRLS. Another explanation could be that the culture of feedback in the classroom varies across countries (and/or successive student cohorts within countries). For example, countries might differ in the extent to which they emphasize social comparisons versus individualized performance feedback (Lüdtke & Köller, 2002; Lüdtke et al., 2005) or in how common it is to apply differentiated instruction (i.e., individually altering students' expectations and goals, varying

the complexity of tasks; Roy et al., 2015). In countries in which teachers provide individualized performance feedback and differentiated instruction, students' achievement and their corresponding self-concepts should be less strongly related than in countries in which teachers foster social comparisons and no differentiated instruction. Until now, little has been known about these cross-cultural variations (or variations within countries that differently affect successive student cohorts), and thus, this topic is an area for future research.

Conclusion

The present study makes three main contributions. First, given the size and representativeness of the applied data and our rigorous integrative data analysis, our results provide strong support for the generalizability of nonlinear relations between achievement and corresponding self-concepts in mathematics and the verbal domain across 13 countries at the secondary school level. Second, in light of the current focus on the replication and reproducibility of research results (e.g., Ioannidis, 2005), we used several PISA, TIMSS, and PIRLS cycles to replicate our findings. This is of particular importance at this point in time, given the lack of findings that have been replicated in many areas of psychology, educational science, and other fields. Third, the present study advances the current theory in self-concept research. Nonlinear effects between achievement and domain-specific self-concepts have mostly been neglected in previous research (but see Marsh, 2004; Marsh & Rowe, 1996). From a theoretical point of view, the results of the present investigation suggest that models on self-concept formation may be refined by integrating nonlinear effects (e.g., quadratic effects) to better approximate empirical relations between achievement and corresponding self-concepts for all students.

References

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Alicke, M. D., & Sedikides, C. (2011). *Handbook of self-enhancement and self-protection.* The Guilford Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

BIFIE (2018). BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 3.0-14. https://CRAN.R-project.org/package=BIFIEsurvey

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528–541. https://doi.org/10.1037/met0000016

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons.

Buckley, J. (2009). *Cross-national response styles in international educational assessments: Evidence from PISA 2006.* https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation.* American Psychological Association.

Byrne, B. M. (2002). Validating the measurement and structure of self-concept: Snapshots of past, present, and future research. *American Psychologist, 57*(11)*,* 897–909. https://doi.org/10.1037/0003-066X.57.11.897

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*(3), 401–423. https://doi.org/10.1037/0033-2909.115.3.401

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*(1521). https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L. & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*(738). https://doi.org/10.3389/fpsyg.2016.00738

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129. https://doi.org/10.2307/3001666

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum.

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods, 14*(2), 77– 80. https://doi.org/10.1037/a0015972

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Eccles, J. S., Midgley, C., & Adler, T. F. (1984). Grade-related changes in the school environment: Effects on achievement motivation. In J. G. Nicholls (Ed.), *The development of achievement motivation* (pp. 283–331). JAI Press.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Eddington, A. (1939). *The philosophy of physical science.* Macmillan.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Frey, K. S., & Ruble, D. N. (1990). Strategies for comparative evaluation: Maintaining a sense of competence across the life span. In R. J. Sternberg & J. Kolligian Jr. (Eds.), *Competence considered* (pp. 167–189). Yale University Press.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*(2), 254–267. https://doi.org/10.1080/14792779143000033

Gogol, K., Brunner, M., Martin, R., Preckel, F., & Goetz, T. (2017). Affect and motivation within and between school subjects: Development and validation of an integrative structural model of academic self-concept, interest, and anxiety. *Contemporary Educational Psychology, 49*, 46–65. https://doi.org/10.1016/j.cedpsych.2016.11.003

Graham, E. K., Rutsohn, J. P., Turiano, N. A., Bendayan, R., Batterham, P. J., Gerstorf, D., Katz, M. J., Reynolds, C. A., Sharp, E. S., Yoneda, T. B., Bastarache, E. D., Elleman, L. G., Zelinski, E. M., Johansson, B., Kuh, D., Barnes, L. L., Bennett, D. A., Deeg, D. J. H., Lipton, R. B., ... & Mroczek, D. K. (2017). Personality predicts mortality risk: An integrative data analysis of 15 international longitudinal studies. *Journal of Research in Personality, 70,* 174–186. https://doi.org/10.1016/j.jrp.2017.07.005

Gramzow, R. H., Elliot, A. J., Asher, E., & McGregor, H. A. (2003). Self-evaluation bias and academic performance: Some ways and some reasons why. *Journal of Research in Personality, 37*(2)*,* 41–61. https://doi.org/10.1016/S0092-6566(02)00535-4

Green, J. D., & Sedikides, C. (2004). Retrieval selectivity in the processing of self-referent information: Testing the boundaries of self-protection. *Self and Identity, 3*(1), 69–80. https://doi.org/10.1080/13576500342000059

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*(1), 124–136. https://doi.org/10.1037/0022-0663.95.1.124

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037//0022-0663.92.1.160

Harter, S. (2012). *The construction of the self: Developmental and sociocultural foundations.* Guilford Press.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*(3), 319–334. https://doi.org/10.1177/0022022116687395

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Hepper, E. G., Gramzow, R. H., & Sedikides, C. (2010). Individual differences in self-enhancement and self-protection strategies: An integrative analysis. *Journal of Personality, 78*(2), 781–814. https://doi.org/10.1111/j.1467-6494.2010.00633.x

Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011). The Cochrane Collaboration. https://www.handbook.cochrane.org

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558. https://doi.org/10.1002/sim.1186

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*(2), 150–164. https://doi.org/10.1037/a0015566

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

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Huguet, P., Dumas, F., Marsh, H., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Regner, 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

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2*(2), 324–347. https://doi.org/10.1037/1076-8971.2.2.324

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

Krueger, J. I., & Wright, J. C. (2011). Measurement of self-enhancement (and self-protection). In M. D. Alicke & C. Sedikides, *Handbook of self-enhancement and self-protection* (pp. 472–494). The Guilford Press.

Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–286). CRC Press.

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, *58*, 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Lüdtke, O., & Köller, O. (2002). Individuelle Bezugsnormorientierung und soziale Vergleiche im Mathematikunterricht Einfluss unterschiedlicher Referenzrahmen auf das fachspezifische Selbstkonzept der Begabung [Individual reference norm and social comparisons in mathematics classes: The impact of different frames of reference on the domain-specific self-concept of ability]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 34*(3), 156–166. https://doi.org/10.1026//0049-8637.34.3.156

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

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software 9*(1), 1–19. https://doi.org/10.18637/jss.v009.i08

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. https://doi.org/10.1037//1082-989X.7.1.19

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/10.2307/1163048

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. (1990a). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology*, *82*(4), 646–656. https://doi.org/10.1037/0022-0663.82.4.646

Marsh, H. W. (1990b). *Self-Description Questionnaire, II.* Psychological Corporation.

Marsh, H. W. (2004). Negative effects of school-average achievement on academic self-concept: A comparison of the big-fish-little-pond effect across Australian states and territories. *Australian Journal of Education, 48*(1)*,* 5–26. https://doi.org/10.1177/000494410404800102

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. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96*(1), 56–67. https://doi.org/10.1037/0022-0663.96.1.56

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., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality, 74*(2)*,* 403–456. https://doi.org/10.1111/j.1467-6494.2005.00380.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

Martin, M.O., & Mullis, I.V.S. (2012). *Methods and procedures in TIMSS and PIRLS 2011.* TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015.* TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016.* TIMSS & PIRLS International Study Center, Boston College.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806– 834. https://doi.org/10.1037/0022-006X.46.4.806

Möller, J., & Pohlmann, B. (2010). Achievement differences and self-concept differences: Stronger associations for above or below average students? *British Journal of Educational Psychology, 80*(3), 435–450. https://doi.org/10.1348/000709909X485234

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*(3)*,* 1129–1167. https://doi.org/10.3102/0034654309337522

Möller, J., Zimmermann, F., & Köller, O. (2014). The reciprocal internal/external frame of reference model using grades and test scores. *British Journal of Educational Psychology, 84*(4), 591–611. https://doi.org/10.1111/bjep.12047

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 assessment framework* (2nd ed.). TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Martin, M. O. (2015). *PIRLS 2016 assessment framework* (2nd ed.). TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L, & Sainsbury, M. (2009). *PIRLS 2011 assessment framework.* TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks.* TIMSS & PIRLS International Study Center, Boston College.

OECD (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy.* OECD Publishing.

OECD (2002). *PISA 2000 technical report.* OECD Publishing.

OECD (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills.* OECD Publishing.

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing.

OECD (2014). *PISA 2012 technical report.* OECD Publishing.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Prast, E. J., Van de Weijer-Bergsma, E., Miočević, M., Kroesbergen, E. H., & Van Luit, J. E. (2018). Relations between mathematics achievement and motivation in students of diverse achievement levels. *Contemporary Educational Psychology, 55,* 84–96. https://doi.org/10.1016/j.cedpsych.2018.08.002

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2)*,* 178–192. https://doi.org/10.1037/1082-989X.10.2.178

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Robitzsch, A., Grund, S., & Henke, T. (2018). *miceadds: Some additional multiple imputation functions, especially for mice. R package version 3.0-16.* https://CRAN.R-project.org/package=miceadds

Roy, A., Guay, F., & Valois, P. (2015). The big-fish–little-pond effect on academic self-concept: The moderating role of differentiated instruction and individual achievement. *Learning and Individual Differences, 42*, 110–116. https://doi.org/10.1016/j.lindif.2015.07.009

Ruble, D. N., & Frey, K. S. (1991). Changing patterns of comparative behavior as skills are acquired: A functional model of self-evaluation. In J. Suls & T. A. Wills (Eds.), *Social comparison: Contemporary theory and research* (pp. 70–112). Erlbaum.

Schurtz, I. M., Pfost, M., & Artelt, C. (2014). Variieren die Selbstkonzeptdifferenzen in Abhängigkeit vom Leistungsniveau? Differenzielle Zusammenhänge in Deutsch, Englisch und Mathematik [Do self-concept differences vary in dependence of the achievement level? Differential relations in language arts, English, and mathematics]. *Zeitschrift für Pädagogische Psychologie, 28*(1-2), 31–42. https://doi.org/10.1024/1010-0652/a000122

Segall, M. H., & Lonner, W. J. (1998). Cross-cultural psychology as a scholarly discipline. On the flowering of culture in behavioral research. *American Psychologist*, *53(10)*, 1101–1110. https://doi.org/10.1037/0003-066X.53.10.1101

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/10.3102/00346543046003407

Simonsohn, U. (2018). Two Lines: A valid alternative to the invalid testing of u-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555. https://doi.org/10.1177/2515245918805755

Skaalvik, E. M., & Rankin, R. J. (1992). Math and verbal achievement and self-concepts: Testing the internal/external frame of reference model. *The Journal of Early Adolescence, 12*(3), 267–279. https://doi.org/10.1177/0272431692012003003

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50*(3–4), 241–252. https://doi.org/10.1023/B:SERS.0000015555.40976.e6

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science, 1*(1), 115–120. https://doi.org/10.1177/2515245918757689

Stipek, D. J., & Daniels, D. H. (1988). Declining perceptions of competence: A consequence of changes in the child or in the educational environment? *Journal of Educational Psychology, 80*(3), 352–356. https://doi.org/10.1037/0022-0663.80.3.352

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science, 11*(5), 159–163. https://doi.org/10.1111/1467-8721.00191

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole Publishing Co.

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: Theory, research, and practice* (pp. 187–214). Springer. https://doi.org/10.1007/978-3-319-28606-8_8

Van der Beek, J. P., Van der Ven, S. H., Kroesbergen, E. H., & Leseman, P. P. (2017). Self-concept mediates the relation between achievement and emotions in mathematics. *British Journal of Educational Psychology, 87*(3), 478–495. https://doi.org/10.1111/bjep.12160

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(9). https://doi.org/10.1186/s40536-014-0009-0

Wheeler, L., & Suls, J. (2005). Social comparison and self-evaluations of competence. In A. J. Elliot & C. S. Dweck, *Handbook of competence and motivation* (pp. 566–578). Guilford Press.

Wickham, H. (2009). *Elegant graphics for data analysis (ggplot2)*. Springer.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A.

Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). Academic Press.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Table 1. *Number of Examined Countries, Sample Sizes, Percentage of Female Students (%F), Number of Participating Classes or Schools, and Students' Mean Age for Individual Assessments and in Total*

Assessment	$N_{countries}$	N	$\%_F$	$N_{classes/schools}$	Age_{mean}
Elementary school students					
TIMSS/PIRLS 2011 (Grade 4)	13[a]	56,868	49	3,159	10.30
TIMSS 2015 (Grade 4)	12[b]	55,642	49	3,104	10.38
PIRLS 2016 (Grade 4)	13[a]	59,586	49	3,305	10.36
Secondary school students					
TIMSS 2011 (Grade 8)	8[c]	39,317	49	1,990	14.35
TIMSS 2015 (Grade 8)	8[d]	42,120	50	2,108	14.42
PISA 2000 (15-year-olds)	13[a]	35,024	51	2,412	15.71
PISA 2003 (15-year-olds)	13[a]	77,910	50	2,868	15.78
PISA 2012 (15-year-olds)	13[a]	104,337	49	4,361	15.76
Total	13[a]	470,804	50	23,307	

Note. In TIMSS and PIRLS, one or more intact classes of students were randomly sampled within randomly selected schools, whereas in PISA, students were randomly sampled within randomly selected schools.

[a] Australia, Austria, Czech Republic, Finland, Germany, Hong Kong, Hungary, Ireland, Italy, Norway, Portugal, Russian Federation, Sweden.

[b] Austria did not participate.

[c] Australia, Czech Republic, Finland, Germany, Hong Kong, Hungary, Ireland, Italy, Norway, Portugal, Russian Federation, Sweden.

[d] Australia, Hong Kong, Hungary, Ireland, Italy, Norway, Russian Federation, Sweden.

Table 2. *Median Reliabilities (and Range) for Mathematics and Reading Achievement and Self-Concept Scales across the Examined Countries*

Assessment	Achievement[a]				Self-concept[b]			
	Mathematics		Reading		Mathematics		Verbal/reading	
Elementary school students								
TIMSS/PIRLS 2011(Grade 4)	.82	(.79-.87)	.87	(.81-.89)	.87	(.85-.90)	.72	(.64-.77)
TIMSS 2015 (Grade 4)	.83	(.81-.88)		–	.87	(.84-.89)		–
PIRLS 2016 (Grade 4)		–	.88	(.85-.91)		–	.80	(.72-.82)
Secondary school students								
TIMSS 2011 (Grade 8)	.89	(.83-.91)		–	.91	(.90-.93)		–
TIMSS 2015 (Grade 8)	.89	(.83-.91)		–	.91	(.89-.92)		–
PISA 2000 (15-year-olds)	.81		.89		.87	(.84-.93)	.76	(.66-.82)
PISA 2003 (15-year-olds)	.91	(.88-.93)		–	.89	(.81-.92)		–
PISA 2012 (15-year-olds)	.93	(.91-.94)		–	.89	(.82-.92)		–

Note. In PISA 2000, only averaged reliabilities were reported for the international PISA scales.

[a] Reliability estimates were based on item response theory.

[b] Reliability was measured as Cronbach's alpha.

Table 3. *Meta-Analytic Results for Mathematics: Average Regression Coefficients and Corresponding Heterogeneity Measures (with 95% Confidence Intervals) from the Linear and Quadratic Regression Models for TIMSS and PISA*

	Linear model			Quadratic model				
	Linear term			Linear term			Quadratic term	
	Estimate		95% CI	Estimate		95% CI	Estimate		95% CI
Elementary school students								
TIMSS 2011 & 2015 (Grade 4)								
β_{mean}	0.38		[0.35, 0.42]	0.39		[0.36, 0.42]	0.04		[0.03, 0.06]
Q	567.092	***		635.405	***		79.804	***	
I^2_{total}	95.45	†††		95.80	†††		68.37	††	
I^2_{Level2}	95.45	†††		95.80	†††		62.10	††	
I^2_{Level3}	0			0			6.27		
τ_{total}	0.08			0.08			0.03		
τ_{Level2}	0.08			0.08			0.03		
τ_{Level3}	0			0			0.01		
k	25			25			25		
Secondary school students								
TIMSS 2011 & 2015 (Grade 8)								
β_{mean}	0.53		[0.48, 0.58]	0.56		[0.52, 0.61]	0.12		[0.08, 0.15]
Q	374.899	***		275.647	***		97.126	***	
I^2_{total}	96.24	†††		94.69	†††		86.04	†††	
I^2_{Level2}	0			0			20.99		
I^2_{Level3}	96.24	†††		94.69	†††		65.05	††	
τ_{total}	0.08			0.06			0.05		
τ_{Level2}	0			0			0.03		
τ_{Level3}	0.08			0.06			0.05		
k	16			16			16		

(table continues)

Table 3 *(Continued)*

	Linear model			Quadratic model				
	Linear term			Linear term			Quadratic term	
	Estimate		95% CI	Estimate		95% CI	Estimate	95% CI
PISA 2000, 2003, & 2012 (15-year-olds)								
β_{mean}	0.37		[0.32, 0.42]	0.38		[0.33, 0.43]	0.12	[0.10, 0.14]
Q	1627.021	***		1573.419	***		154.829	***
I^2_{total}	97.63	†††		97.60	†††		75.82	†††
I^2_{Level2}	46.65	†		42.60	†		18.44	
I^2_{Level3}	50.99	††		55.00	††		57.38	††
τ_{total}	0.11			0.10			0.03	
τ_{Level2}	0.07			0.07			0.02	
τ_{Level3}	0.08			0.08			0.03	
k	39			39			39	

Note. β_{mean} = Average regression coefficient, Q = Cochran's measure of homogeneity (Cochran, 1954), I^2_{total} = Higgins and Thompson's (2002) measure of (total) heterogeneity, I^2_{Level2} = Percentage of the variability in regression coefficients that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in regression coefficients that is due to heterogeneity between countries rather than sampling error, τ_{total} = Total standard deviation of regression coefficients, τ_{Level2} = Within-country standard deviation of regression coefficients, τ_{Level3} = Between-country standard deviation of regression coefficients, k = number of countries.

† Moderate heterogeneity (I^2 > 30%), †† Substantial heterogeneity (I^2 > 50%), ††† Considerable heterogeneity (I^2 > 75%).

*** p < .001.

Table 4. *Meta-Analytic Results for the Verbal Domain: Average Regression Coefficients and Corresponding Heterogeneity Measures (with 95% Confidence Intervals) from the Linear and Quadratic Regression Models for PIRLS and PISA*

	Linear model			Quadratic model					
	Linear term			Linear term			Quadratic term		
	Estimate		95% CI	Estimate		95% CI	Estimate		95% CI
Elementary school students									
PIRLS 2011 & 2016 (Grade 4)									
β_{mean}	0.44		[0.41, 0.46]	0.43		[0.41, 0.46]	-0.02		[-0.03, -0.01]
Q	248.447	***		244.134	***		57.989	***	
I^2_{total}	88.41	†††		88.62	†††		54.48	††	
I^2_{Level2}	14.29			12.34			46.42	†	
I^2_{Level3}	74.12	††		76.28	†††		8.06		
τ_{total}	0.05			0.05			0.02		
τ_{Level2}	0.02			0.02			0.02		
τ_{Level3}	0.04			0.04			0.01		
k	26			26			26		
Secondary school students									
PISA 2000 (15-year-olds)									
β_{mean}	0.27		[0.23, 0.31]	0.28		[0.24, 0.32]	0.05		[0.03, 0.07]
Q	154.299	***		169.885	***		28.171	**	
I^2_{total}	91.99	†††		92.67	†††		53.65	††	
τ_{total}	0.07			0.08			0.03		
k	13			13			13		

Note. β_{mean} = Average regression coefficient, Q = Cochran's measure of homogeneity (Cochran, 1954), I^2_{total} = Higgins and Thompson's (2002) measure of (total) heterogeneity, I^2_{Level2} = Percentage of the variability in regression coefficients that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in regression coefficients that is due to heterogeneity between countries rather than sampling error, τ_{total} = Total standard deviation of regression coefficients, τ_{Level2} = Within-country standard deviation of regression coefficients, τ_{Level3} = Between-country standard deviation of regression coefficients, k = number of countries.

† Moderate heterogeneity (I^2 > 30%), †† Substantial heterogeneity (I^2 > 50%), ††† Considerable heterogeneity (I^2 > 75%).
** p < .01. *** p < .001.

Table 5. *Meta-Analytic Results for Mathematics: Average Regression Coefficients and Corresponding Heterogeneity Measures (with 95% Confidence Intervals) from the Interrupted Regressions for TIMSS and PISA*

	β_1			β_2		
	Estimate		95% CI	Estimate		95% CI
Elementary school						
TIMSS 2011 & 2015 (Grade 4)						
β_{mean}	0.32		[0.26, 0.39]	0.34		[0.29, 0.38]
Q	160.911	***		37.352	*	
I^2_{total}	92.14	†††		41.16	†	
I^2_{Level2}	92.14	†††		41.16	†	
I^2_{Level3}	0			0		
τ_{total}	0.14			0.05		
τ_{Level2}	0.14			0.05		
τ_{Level3}	0			0		
k	25			25		
Secondary school						
TIMSS 2011 & 2015 (Grade 8)						
β_{mean}	0.25		[0.08, 0.42]	0.58		[0.50, 0.66]
Q	294.514	***		107.031	***	
I^2_{total}	98.24	†††		89.41	†††	
I^2_{Level2}	46.29	†		67.39	††	
I^2_{Level3}	51.94	††		22.03		
τ_{total}	0.28			0.12		
τ_{Level2}	0.19			0.11		
τ_{Level3}	0.20			0.06		
k	16			16		
PISA 2000, 2003, & 2012 (15-year-olds)						
β_{mean}	0.03		[-0.01, 0.07]	0.47		[0.41, 0.53]
Q	84.347	***		1281.493	***	
I^2_{total}	52.68	††		97.22	†††	
I^2_{Level2}	49.29	†		22.84		
I^2_{Level3}	3.40			74.39	††	
τ_{total}	0.08			0.12		
τ_{Level2}	0.08			0.06		
τ_{Level3}	0.02			0.10		
k	39			39		

Note. β_{mean} = Average regression coefficient, Q = Cochran's measure of homogeneity (Cochran, 1954), I^2_{total} = Higgins and Thompson's (2002) measure of (total) heterogeneity, I^2_{Level2} = Percentage of the variability in regression coefficients that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in regression coefficients that is due to heterogeneity between countries rather than sampling error, τ_{total} = Total standard deviation of regression coefficients, τ_{Level2} = Within-country standard deviation of regression coefficients, τ_{Level3} = Between-country standard deviation of regression coefficients, k = number of countries.

† Moderate heterogeneity ($I^2 > 30\%$), †† Substantial heterogeneity ($I^2 > 50\%$), ††† Considerable heterogeneity ($I^2 > 75\%$).

* $p < .05$. *** $p < .001$.

Table 6. *Meta-Analytic Results for the Verbal Domain: Average Regression Coefficients and Corresponding Heterogeneity Measures (with 95% Confidence Intervals) from the Interrupted Regressions for PIRLS and PISA*

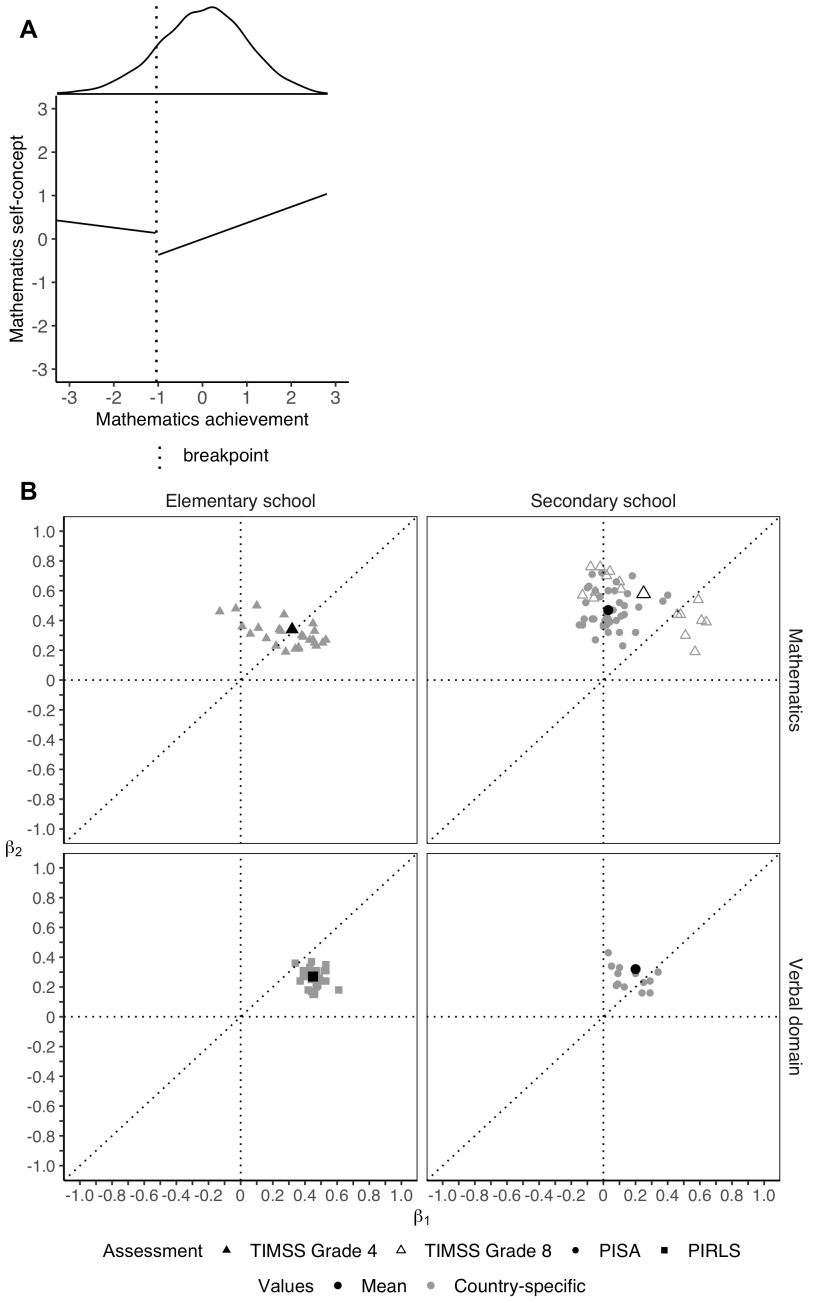
	β_1		β_2	
	Estimate	95% CI	Estimate	95% CI
Elementary school				
PIRLS 2011 & 2016 (Grade 4)				
β_{mean}	0.45	[0.42, 0.48]	0.27	[0.22, 0.31]
Q	130.861 ***		6.737	
I^2_{total}	77.01 †††		0	
I^2_{Level2}	11.84		0	
I^2_{Level3}	65.16 ††		0	
τ_{total}	0.05		0	
τ_{Level2}	0.02		0	
τ_{Level3}	0.05		0	
k	26		26	
Secondary school				
PISA 2000 (15-year-olds)				
β_{mean}	0.20	[0.14, 0.26]	0.32	[0.25, 0.39]
Q	40.103 ***		23.551 *	
I^2_{total}	84.03 †††		51.64 ††	
τ_{total}	0.08		0.05	
k	13		13	

Note. β_{mean} = Average regression coefficient, Q = Cochran's measure of homogeneity (Cochran, 1954), I^2_{total} = Higgins and Thompson's (2002) measure of (total) heterogeneity, I^2_{Level2} = Percentage of the variability in regression coefficients that is due to heterogeneity within countries rather than sampling error; I^2_{Level3} = Percentage of the variability in regression coefficients that is due to heterogeneity between countries rather than sampling error, τ_{total} = Total standard deviation of regression coefficients, τ_{Level2} = Within-country standard deviation of regression coefficients, τ_{Level3} = Between-country standard deviation of regression coefficients, k = number of countries.

†† Substantial heterogeneity ($I^2 > 50\%$), ††† Considerable heterogeneity ($I^2 > 75\%$).

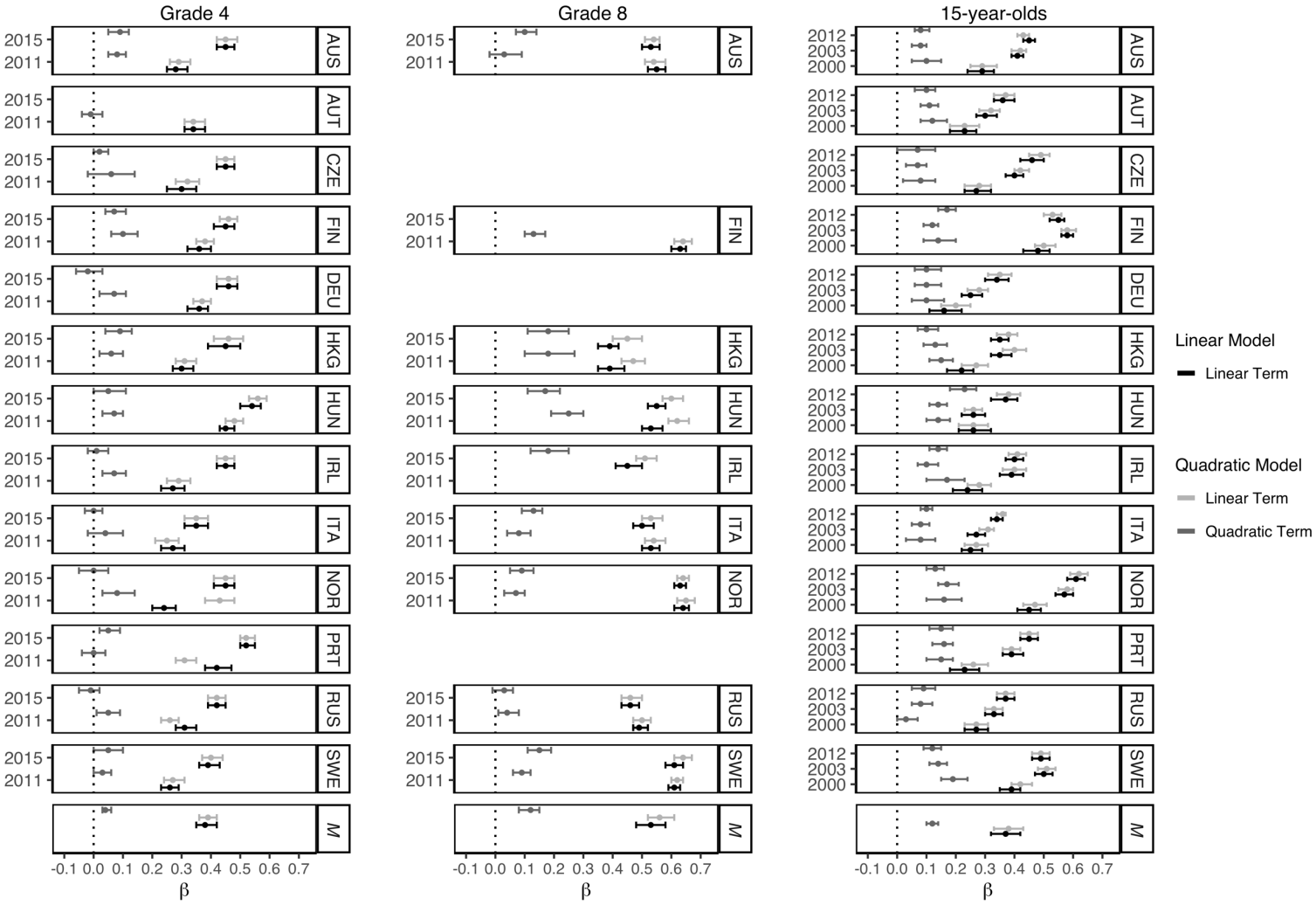
* $p < .05$. *** $p < .001$.

Figure 1. *Domain-Specific Pattern of Slopes as Obtained from the Interrupted Regressions: (A) Example Graph for a Single Country (Australia in PISA 2000)*



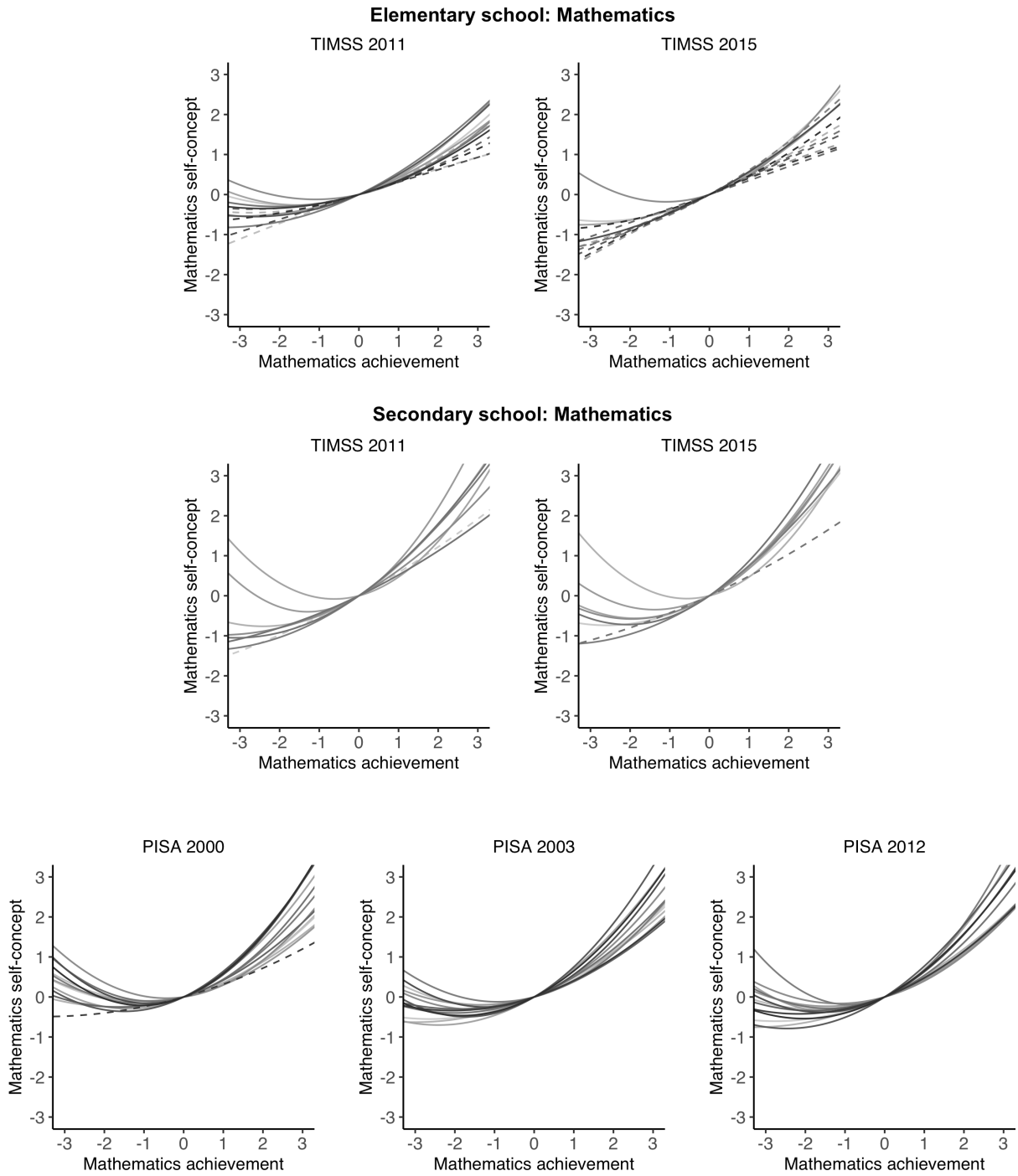
Note. The density plot shows the math achievement distribution for the first plausible value. (B) Summary of results depicted separately for mathematics (TIMSS and PISA) and the verbal domain (PIRLS and PISA) and for elementary school students (TIMSS and PIRLS) and secondary school students (TIMSS and PISA). β_1 represents the regression coefficient to the left of the breakpoint (i.e., relatively lower achieving students), β_2 the regression coefficient to the right of the breakpoint (i.e., relatively higher achieving students). Symbols above the diagonal indicate that the relation between achievement and self-concept was weaker for lower achieving students than for higher achieving students in a certain country or assessment. Mean values are depicted in black, country-specific values in grey.

Figure 2. *Country-Specific Relations Between Mathematics Achievement and Mathematics Self-Concept for Different Age Groups (Left: TIMSS 2011/2015 Grade 4, Middle: TIMSS 2011/2015 Grade 8, Right: PISA 2000/2003/2012)*



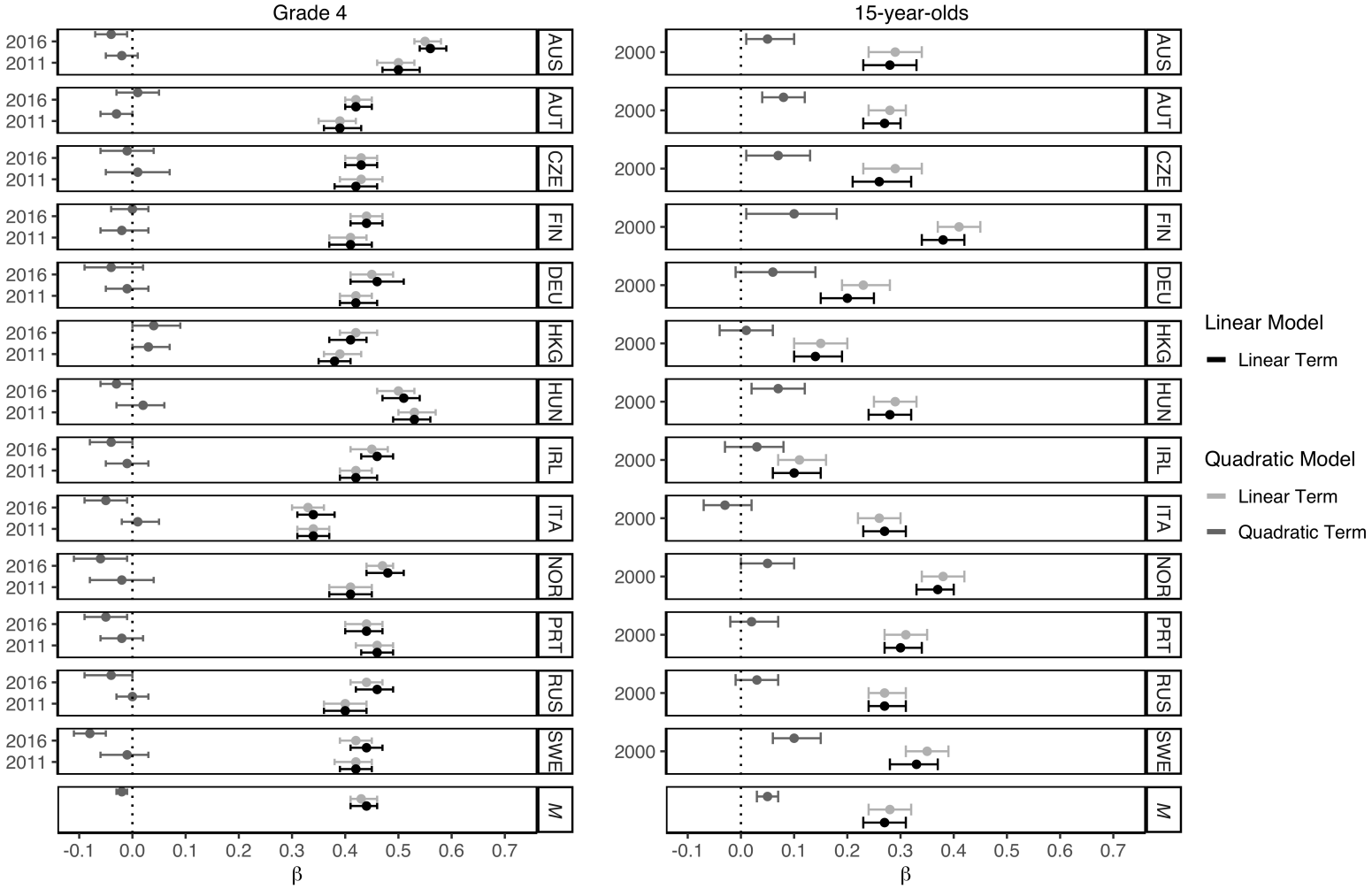
Note. AUS = Australia, AUT = Austria, CZE = Czech Republic, FIN = Finland, DEU = Germany, HKG = Hong Kong, HUN = Hungary, IRL = Ireland, ITA = Italy, NOR = Norway, PRT = Portugal, RUS = Russian Federation, SWE = Sweden, *M* = weighted mean.

Figure 3. *Country-specific Quadratic Regression Lines for the Mathematics Domain (TIMSS and PISA Assessments)*



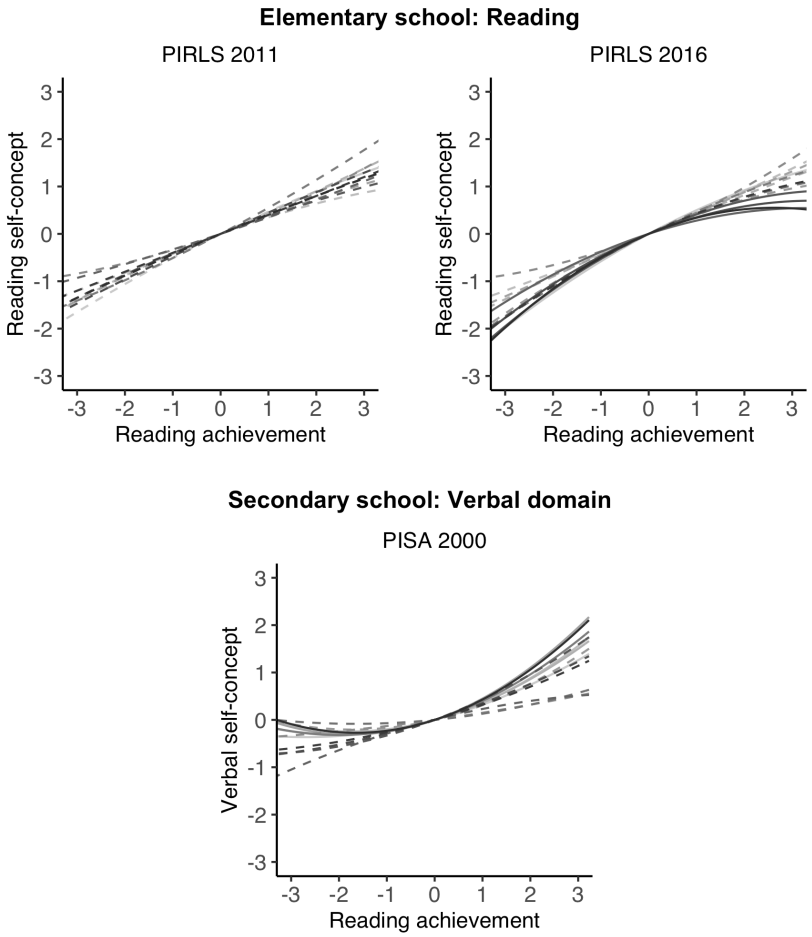
Note. Significant quadratic effects are depicted as solid lines, nonsignificant quadratic effects as dashed lines.

Figure 4. *Country-Specific Relations Between Reading Achievement and Reading Self-Concept (Left: PIRLS 2011/2016) or Verbal Self-Concept (Right: PISA 2000)*



Note. AUS = Australia, AUT = Austria, CZE = Czech Republic, FIN = Finland, DEU = Germany, HKG = Hong Kong, HUN = Hungary, IRL = Ireland, ITA = Italy, NOR = Norway, PRT = Portugal, RUS = Russian Federation, SWE = Sweden, *M* = weighted mean.

Figure 5. Country-Specific Quadratic Regression Lines for the Verbal Domain (PIRLS and PISA Assessments)



Note. Significant quadratic effects are depicted as solid lines, nonsignificant quadratic effects as dashed lines.

4

General Discussion

4 General Discussion

The aim of this doctoral thesis was to study the interplay between achievement and achievement motivation within the Situated Expectancy–Value Theory (SEVT) framework. In two empirical studies, I examined gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in mathematics, reading, and science and their relation to gender equality as well as the functional relation between academic achievement and corresponding self-concepts. A major strength of this dissertation is that the results of both studies are highly robust and generalizable due to the use of research synthesis methods (i.e., performing a multilevel integrative data analysis and a multilevel meta-analysis) and representative individual student data from international large-scale assessments (PISA, TIMSS, PIRLS). Based on the results of these studies, I will answer the research questions of this doctoral thesis:

(1) *What is the extent of gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in mathematics, reading, and science across countries?*

(2) *To what extent are cross-national gender differences in the group of top-performing math students related to sociocultural factors, or more specifically, to the level of gender equality in a country?*

(3) *Which functional relation exists between students' academic achievement and corresponding academic self-concepts?*

In the following sections, I will briefly summarize und discuss the main findings from the two studies along these research questions on the interplay between achievement

and achievement motivation. Subsequently, I will outline general limitations as well as directions for future research and practice before concluding with the final remarks.

4.1 Research Question I: What Is the Extent of Gender Differences in Top-Performing Math Students Achievement, Achievement Profiles, and Achievement Motivation Across Countries?

Study I meta-analyzed gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation in mathematics, reading, and science. To capture the interplay of achievement and achievement motivation, achievement, achievement profiles, and achievement motivation were analyzed together in the population of 15-year-old students who scored in the top 5% in mathematics. Capitalizing on data from six PISA cycles (PISA 2000–2015), we calculated 55 standardized mean effect sizes for gender differences across 82 countries.

4.1.1 Results From Study 1

Our results showed that there were on average more male than female students (40%) that scored in the top 5% in mathematics. Furthermore, we found moderate gender differences in top-performing math students' interest in specific science topics and their verbal motivation. Male students reported higher interest in physics-related topics than female students (i.e., physics, motion of forces, energy transformation), whereas female students were more interested in human biology than their male counterparts. Furthermore, female students reported a higher interest in reading, they enjoyed reading more, and had a higher verbal self-concept than male students. Importantly, mathematically top-performing male students' achievement profiles were on average more mathematics-oriented; on the contrary, mathematically top-performing female students' achievement profiles were more

balanced across domains. Gender differences in top-performing math students' achievement in mathematics, reading, and science as well as in their mathematics motivation and in most aspects of their science motivation were negligible to small. Overall, the pattern of gender differences in the group of top-performing math students was similar to the pattern of gender differences found for the general population (reviewed in Chapter 1; for a systematic comparison, see Table 1).

Gender differences in the group of top-performing math students are important antecedents of gender differences in STEM choices (e.g., Ceci et al., 2014; Eccles, 1994). Study I provided strong evidence that there were (1) less female students in the group of top-performing math students and (2) important gender differences in top-performing math students' specific science interests, motivation in the verbal domain, and achievement profiles. This overall pattern of more balanced achievement profiles in mathematics, reading, and science as well as a stronger verbal motivation might give the average mathematically talented female student broader career options than the average mathematically talented male student, which might contribute to women's underrepresentation in STEM.

Table 1. *Comparison of Proportions of Effect Sizes (in Percent) for Gender Differences Achievement, Achievement Motivation, and Profile Nonoverlap in Mathematics, Reading, and Science That Were Negligible, Small, Moderate, Large, or Very Large in the Group of Top-Performing Math Students' (Study I) and in the General Population*

	Achievement						Achievement motivation						Profile nonoverlap		
	Math		Reading		Science		Math		Reading		Science		M–R	S–R	M–S
Magnitude	Top 5[a]	GP[b]	Top 5[c]	GP[d]	Top 5[a]	GP[e]	Top 5[f]	GP[g]	Top 5[h]	GP[i]	Top 5[j]	GP[k]	Top 5[c]	Top 5[c]	Top 5[a]
Negligible	x	52		38	x	23	22	22	0	0	44	35			
Small		48	x	57		46	78	72	0	50	33	27			x
Moderate		0		5		31	0	6	100	33	22	23			
Large		0		0		0	0	0	0	17	0	8	x	x	
Very large		0		0		0	0	0	0	0	0	8			

Note. Top 5 = Students in the top 5% in mathematics; GP = General population; M–R = Math–reading profile; S–R = Science–reading profile; M–S = Math–science profile; x = Only one effect size available. Figures may not add up to 100% due to rounding. For achievement and achievement motivation: Negligible = $0.00 \leq |d| \leq 0.10$, small = $0.10 < |d| \leq 0.35$, moderate = $0.35 < |d| \leq 0.65$, large = $0.65 < |d| \leq 1.00$, very large = $|d| > 1.00$. k = Number of effect sizes, n_{TOP5} = Number of countries, n_{GP} = Number of studies. For profile nonoverlap: Negligible = 0 % ≤ nonoverlap ≤ 8%, small = 8% < nonoverlap ≤ 24%, moderate = 24% < nonoverlap ≤ 41%, large = 41% < nonoverlap ≤ 55%, very large = nonoverlap > 55%.

[a] $k = 343$, $n_{TOP5} = 82$
[b] $k = 1905$, $n_{GP} = 13$
[c] $k = 342$, $n_{TOP5} = 82$
[d] $k = 1008$, $n_{GP} = 12$
[e] $k = 1264$, $n_{GP} = 7$
[f] $k = 875$, $n_{TOP5} = 65$
[g] $k = 1258$, $n_{GP} = 10$
[h] $k = 180$, $n_{TOP5} = 73$
[i] $k = 201$, $n_{GP} = 5$
[j] $k = 1207$, $n_{TOP5} = 72$
[k] $k = 847$, $n_{GP} = 7$

4.1.2 Gender Differences in Interests in Specific Science Topics: Are Female Students Just Not Interested in These Areas?

Among the largest gender differences we found in Study I involved that mathematically top-performing male student had greater interest in physics-related topics and mathematically top-performing female students had greater interest in human biology. This results pattern is consistent with previous meta-analyses that examined occupational interests in adults in the general population (Su et al., 2009; Su & Rounds, 2015). Can we thus conclude that female students who excel in mathematics are simply less interested in physics or engineering? Should society accept it as a fact and not encourage female students to pursue these domains? On the contrary, it should be investigated how and why these gender differences in interests in specific science topics develop.

Research shows that interest in STEM is not only stimulated through school education, but also through informal contact in extracurricular activities or science competitions (e.g., Dabney et al., 2012; Maltese & Tai, 2010; Sahin et al, 2015; Simpkins et al., 2005). Importantly, female students are less likely to report out-of-school science activities, such as reading about or watching programs about science or extracurricular experiences with batteries, electric toys, fuses, microscopes, or pulleys (Dabney et al., 2012; Jones et al., 2000; Maltese & Tai, 2010). Socialization practices seem to play an important role in whether or to what extent girls and boys are exposed to science activities or tasks early in life. For example, a study that investigated naturally occurring family conversations at a science museum found that parents were three times more likely to explain interactive science exhibits to their sons than to their daughters (aged 1 to 8), although boys and girls were equally interested in the exhibits (Crowley et al., 2001). Furthermore, parents at home and caregivers in kindergarten read books about life science

more often to young boys than to young girls, regardless of the children's interest (Mantzicopoulos & Patrick, 2010). Moreover, parents were more likely to buy math and science toys for boys than for girls in all investigated grade levels (1 to 6; Jacobs & Bleeker, 2004). In addition, during a physics teaching activity, fathers tended to use more scientific vocabulary, asked more conceptual questions, and discussed causal explanations for observations more with sons than with daughters (Tenenbaum & Leaper, 2003). These findings suggest that boys may have greater opportunities than girls to learn about science and practice science skills and consequently to develop interest in specific science topics throughout their childhood. Overall, it is not clear how gender differences in STEM-related interests develop from childhood to young adulthood (e.g., Oppermann et al., 2020). Thus, longitudinal studies that cover the period from kindergarten to high school and beyond are highly desirable.

As discussed in Study I, the answer to the question why mathematically talented male and female students are interested in different STEM domains is likely related to the gender-specific socialization processes. According to the SEVT, they influence students' achievement and achievement motivation and consequently lead to differences between male and female students in educational contexts. As a complement or as an alternative to SEVT, the gender-specific interests in STEM domains may be related to women's and men's goal (communal vs. agentic) orientations that are congruent with their gender roles (e.g., Diekman et al., 2011; Eagly et al., 2020). According to the SRT (Eagly, 1987; Sczesny et al., 2019; Wood & Eagly, 2012) and specifically the related role congruity model (Diekman et al., 2010, 2011), mathematically talented female students might perceive a better match between studying human biology and their communal goals (e.g., to help people) than studying physical and engineering-related sciences. As a result, they select and pursue goals that are congruent to their gender-specific communal roles.

Practical measures to increase women's representation in STEM are discussed in the Practical Implications section.

4.1.3 Comparison of Gender Gaps in Mathematics Achievement

In Study I, we found negligible gender differences in top-performing math students' mathematics achievement ($d = 0.05$). As shown in the Introduction section, there are very few other meta-analyses (or large-scale studies) that focused on gender differences in top-performing math students. Among them, one study investigated gender gaps in mathematics achievement in the top 5% in mathematics using TIMSS and PISA samples and found a small gender gap in favor of male students ($d = 0.15$; Baye & Monseur, 2016). Two other studies meta-analyzed gender gaps in mathematics achievement in academically talented students drawing (at least in part) on data from talent search programs and found moderate gender differences in favor of males ($0.40 \leq d \leq 0.54$; Hyde et al., 1990; Lindberg et al., 2010). Overall, these gender differences were larger than those reported in Study I. There could be two reasons for this, one relating to the samples included (Hyde et al., 1990; Lindberg et al., 2010) and the other to the analytical approach used (Baye & Monseur, 2016).

The gender gap in mathematics achievement in Study I might be smaller than the gender gaps in mathematics achievement reported in the studies by Hyde et al. (1990) and Lindberg et al. (2010) because they included more highly selected samples from talent search programs (e.g., the Study of Mathematically Precocious Youth). Talent search programs that use off-level tests to identify academically talented students (e.g., 12-year-olds are tested with college entrance exams, such as the SAT) show that the variability in test scores is high at the highest level of achievement (e.g., Achter et al., 1996). However, large-scale assessments such as PISA are not designed to differentiate abilities of top-

performers. Thus, in these assessments, gender differences might be somewhat underestimated in top-performing students. However, PISA is considered more challenging than other mathematics assessments (e.g., NAEP or TIMSS; Else-Quest et al., 2010; Hyde et al., 2008). Furthermore, samples from talent search programs have the major disadvantage that they represent a selected student group, particularly in that students are aware of their ability because of their selection into the program. This awareness most likely influences their self-beliefs, motivation, and possibly also their performance. In contrast, data from PISA are representative.

Furthermore, the gender gap in mathematics achievement in Study I might be smaller than the gender gap reported by Baye and Monseur (2016) because of a different analytical approach. Whereas Baye and Monseur (2016) calculated one standardized mean effect size per country and TIMSS and PIRLS cycle and then simply averaged all results to receive one single effect size, we used more sophisticated multilevel random-effects meta-analytic methods. In addition, it is unclear how Baye and Monseur (2016) standardized their effect sizes. If they used the standard deviation of the top 5% for standardization, effect sizes are larger because scores vary less within the top 5% than in the general population. In Study I, we used the standard deviation of the full student population to calculate standardized effect sizes, which is the preferred analytical approach (Cumming & Calin-Jageman, 2016).

4.1.4 How Big Is Small? On the Practice of Benchmarking Effect Sizes

To facilitate the comparison of gender differences in top-performing math students' achievement, achievement motivation, and achievement profiles with the results from other studies, we calculated standardized effect sizes (Cohen's *d*) in Study I. However, because standardized effect sizes are measured on an abstract scale (i.e., standard deviation

units of the outcome measure), their magnitudes are hard to interpret (Baird & Pane, 2019). The standard approach to evaluating the magnitude of effect sizes is to apply effect size benchmarks suggested by Cohen (1969; small = $0.2 < |d| \leq 0.5$, moderate = $0.5 < |d| \leq 0.8$, large = $|d| < 0.8$). Although these benchmarks are widely used, a generalized application is not advisable as they are based on the results of a few social-psychological laboratory studies (Kraft, 2020). Cohen (1988) himself proposed that his benchmarks were "recommended for use only when no better basis for estimating the [effect size] index is available" (p. 25). Therefore, to evaluate the magnitude of gender differences in Study I, we used effect size benchmarks by Hyde (2005) that set thresholds for negligible ($0.00 \leq |d| \leq 0.10$), small ($0.10 < |d| \leq 0.35$), moderate ($0.35 < |d| \leq 0.65$), large ($0.65 < |d| \leq 1.00$), and very large ($|d| > 1.00$) effects. Hyde (2005) derived these benchmarks from the results of an overview of 128 meta-analytical effects representing gender differences in a variety of psychological variables.

Based on these benchmarks, we considered many gender differences in top-performing students' achievement and achievement motivation in Study I to be negligible or small. However, it is important to note that also small effect sizes can have practical importance. Small effects can have considerable influence when gender differences persist across time and situations (Eagly, 2013). As noted by Abelson (1985, p. 133), "small variance contributions of independent variables in single-shot studies grossly understate the variance contribution in the long run" (see Abelson, 1985 and Rosenthal, 1990 for examples of small effects with large practical importance). Especially for intervention studies, efforts have recently been made to improve the evaluation of the practical importance of effects by translating research results into new metrics, such as the percentile growth (i.e., the change in percentile rank that would have been experienced by the median student in the control group, if the student had received the intervention; Baird

& Pane, 2019). In the past, it has been quite popular to express the practical importance of differences in PISA scores in units of years of learning (e.g., if female students outperformed male students in reading by 20 points, this would translate into an advantage in reading for female students of half a year of schooling; Schleicher, 2019). However, the review of Baird and Pane (2019) advises against this translation for several methodological reasons, with an important argument being that the years-of-learning metric ignores the statistical uncertainty of effects even though the translation often substantially increases uncertainty. For example, the authors found that confidence intervals ranged between one quarter of a year and 5,000 years (Baird & Pane, 2019).

Eagly (2013) argues that for the further theoretical development, not the classification into small, medium, and large gender differences is essential, but the investigation and explanation of the variation in gender differences. The following section examines one possible explanation for the cross-national variation of gender differences in top-performing math students.

4.2 Research Question II: To What Extent Are Cross-National Gender Differences in the Group of Top-Performing Math Students Related to the Level of Gender Equality in a Country?

The results of the meta-analysis presented in the preceding section further indicated that many of the gender differences in top-performing math students' achievement, achievement profiles, and achievement motivation varied across countries (Study I). According to the SEVT (e.g., Eccles, 1994) and the SRT (e.g., Wood & Eagly, 2012), gender differences should be smaller in countries with higher levels of gender equality (see Chapter 1 and Study I). Thus, to examine whether the heterogeneity in effect sizes was related to the level of gender equality in a country, we conducted multivariate moderator

analyses. We selected domain-specific gender equality indicators in the areas of education (i.e., gender ratios in primary, secondary, and tertiary education enrollment) and higher positions (i.e., women's share of higher positions and research positions in a country) that are theoretically relevant as moderators of girls' and women's engagement in education and in particular in mathematics (Else-Quest & Grabe, 2012).

4.2.1 Results From Study I

The moderator analyses revealed that a higher share of female students in tertiary education was positively related to the share of female students in the top 5% in mathematics. Moreover, a higher share of female students in tertiary education and a higher share of women in research positions predicted smaller differences in female (and male) students' achievement profile scores. However, gender gaps in top-performing students' achievement and achievement motivation in math, reading, and science were not substantively related to the level of gender equality in a country. Thus, the prediction of the SEVT (e.g., Eccles, 1994) and the SRT (e.g., Wood & Eagly, 2012) that gender differences should be smaller in more gender equal societies was not univocally but at least partially confirmed for the group of top-performing math students.

4.2.2 Alternative Explanations

The fact that cross-national variation in the level of gender equality did not fully explain the variation in gender differences in the group of top-performing math students indicates that there might be further explanatory factors. In the following, I will discuss three possible factors.

As mentioned in Study I, one explanation could be that the effect sizes between and within countries probably varied too little (see Tau values in Tables 3 and 5 in Study I) to

detect moderating relations between the gender equality indicators and gender differences in top-performing math students' achievement and achievement motivation. However, this is true for some, but not all identified gender differences. Hence, another explanation could be that the domain-specific gender equality indicators that we chose were not exhaustive to represent all facets of gender equality. The available gender equality indicators that we selected reflect the gendered access to opportunity structures and the cultural value of women in a society (gender ratios in primary, secondary, and tertiary education enrollment) and the permeation of the so-called glass ceiling (women's share of higher positions and research positions). Thus, they indicate whether it is (a) possible for women to engage in education, and (b) whether it is worth it (Else-Quest & Grabe, 2012). It is highly plausible that variations in these factors contribute to gender differences in educational outcomes across countries.

However, another likely explanation for gender differences in students' domain-specific achievement and achievement motivation could be the varying endorsement of gender stereotypes related to education across countries. Our findings suggest that the average gender differences were consistent with prevailing gender stereotypes (i.e., math + physics = male, verbal domain + human biology = female). Thus, it would be plausible that the perceived importance of specific educational domains (such as mathematics, reading/native language education, and science) for girls/women and boys/men moderated gender differences in these domains. For example, this could be measured by items such as "Education in mathematics/native language/science is more important for a boy than for a girl." Such a measure would assess gendered stereotypes about mathematics, reading, and science at country level and would thus allow to test central predictions of the SEVT and the SRT.

Furthermore, it is possible that other factors than gender equality influence the presence or absence of gender differences in the group of top-performing math students across countries. For instance, studies by Baye and Monseur (2016) and Gray et al. (2019) showed that male students' mathematics achievement scores varied more than female students' scores across a large range of countries. The greater variability of male students' math scores in the general population could be related to the preponderance of male students in the top 5% in mathematics (e.g., Hedges & Friedman, 1993). Reasons for male students' greater variability are mostly unknown (Gray et al., 2019). However, the finding that the variability varied across countries (Baye & Monseur, 2016; Gray et al., 2019) provides first evidence for the influence of further moderating factors. Gray et al. (2019) showed that the variability is to some extent related to the level of gender equality in a country. To this end, a move toward an intersectional perspective that considers the interaction between different social identities (e.g., race, social class, and sexual orientation) simultaneously when studying gender differences will most likely expand our understanding of gender differences—in the general population, but also in the top 5% in mathematics (e.g., Cole, 2009; Hyde, 2014; Parker et al., 2019).

4.2.3 Gender Equality Indicators: Challenges in the Field

Gender equality indicators are an important resource for researchers interested in gender disparities. To improve the status of girls and women across multiple domains, the UN decided in 1995 to expand the data base on women and their status in relation to economic, social, political, cultural, and health-related development (Else-Quest & Hamilton, 2018). Since then, an assortment of composite and domain-specific gender equality indicators have been developed or made accessible (for reviews, see Else-Quest & Grabe, 2012; Else-Quest & Hamilton, 2018; Hawken & Munk, 2013).

Among the most prominent composite indicators are the World Economic Forum's Global Gender Gap Index (GGGI), the UNDP's Gender Empowerment Measure (GEM), the UNDP's Gender Equality Index (GEQ), the UNDP's Gender Inequality Index (GII), Social Watch's Gender Equity Index (GEI), and the OECD's Social Institutions and Gender Index (SIGI). Composite indicators aggregate multiple domains of gender equality (e.g., health, education, politics, and economic participation) into one score. They differ with respect to the domains they include, the domain-specific indicators they choose to represent the domains, and the weighting of the domain-specific indicators to aggregate the composite score (Else-Quest et al., 2010; Hawken & Munk, 2013).

Composite indicators have been and are still widely used to investigate, for example, whether different levels of gender equality across countries moderate gender differences in educational outcomes (e.g., Else-Quest et al., 2010; Gray et al., 2019; Guiso et al., 2008; Hyde & Mertz, 2009; Machin & Pekkarinen, 2008; Reilly, 2012; Stoet & Geary, 2013, 2015, 2018, 2020a; notably, Else-Quest et al., 2010; Reilly, 2012; Reilly et al., 2019; and Stoet & Geary, 2015 also included domain-specific indicators). From a psychometric perspective, however, the use of composite indicators in correlational research designs is highly problematic for several reasons. Most importantly, gender equality is a multidimensional construct. As such, measuring it by one measure cannot achieve any real construct validity (Else-Quest & Hamilton, 2018). A systematic analysis of the measurement methodology of five of the aforementioned composite scores by Hawken and Munk (2013) casts doubts on the psychometric quality of these indicators. For example, the authors found that the considered composite indicators often lack a clear theoretical foundation, including one of the most popular composite gender equality indicators, the GGGI. Furthermore, all indicators share the weaknesses that the set of indicators do not fully cover the meaning of the concept or domain being measured, or

comprise redundant or irrelevant indicators; there was no justification given for various decisions made in the aggregation process; and no sensitivity analyses were carried out to validate the measures (Hawken & Munk, 2013).

Instead, gender equality should be measured by domain-specific indicators that are theoretically relevant and (better) reflect the mechanism under investigation (for examples, see Study I; Else-Quest & Grabe, 2012; Else-Quest & Hamilton, 2018). The advantages of domain-specific indicators are illustrated by the following example: Two nations can achieve the same value on the GGGI, but show variation within and between countries on different domain-specific indicators. I will illustrate this with the example of Argentina and Moldova that both achieve a GGGI score of 0.773 (Table 2). However, Argentina scores relatively high on gender equality in the political domain, but relatively low on gender equality in the economic domain. On the contrary, Moldova scores relatively high on gender equality in the economic domain, but relatively low on gender equality in the political domain. On the between-country level, Argentina scores higher on gender equality in the political domain than Moldova, but lower in the health and economic domains. The example of Argentina and Moldova shows that composite indicators cannot indicate which domains are relevant and, thus, they neither provide empirical evidence for theory development and evaluation, nor for how a country can improve the equality between men and women (Else-Quest & Hamilton, 2018). Domain-specific indicators, by contrast, offer all these possibilities (Else-Quest & Hamilton, 2018) and have been applied in a range of studies to explain which factors moderate cross-national gender gaps in mathematics achievement (Baker & Jones, 1993; Else-Quest et al., 2010; Penner, 2008; Reilly, 2012; Riegel-Crumb, 2005).

Table 2. *Comparison of Domain-Specific Indicators Based on the Example of Two Countries With the Same Value on the GGGI*

Indicators	Argentina	Moldova
GGGI value[a]	0.733	0.733
GGGI rank[a]	36	35
Adolescent birth rate (per 1,000 women ages 15–49 years)[b]	62.80	22.40
Female-to-mal ratio in labor force participation[c]	0.70	0.89
Share of seats in parliament (%)[b]	39.50	22.80

[a] Retrieved from http://www3.weforum.org/docs/WEF_GGGR_2018.pdf
[b] Retrieved from http://hdr.undp.org/en/data
[c] Retrieved from https://data.worldbank.org/indicator/SL.TLF.CACT.FM.ZS

To conclude, due to the low psychometric quality of composite gender equality indicators (Hawken & Munk, 2013), these indicators should not be used to examine the influence of gender equality on any outcomes (Else-Quest & Grabe, 2012; Else-Quest & Hamilton, 2018). Instead, a large number of specific gender equality indicators are available that allow a theory-driven investigation of the relation between specific aspects of gender equality and a specific outcome (see Else-Quest & Grabe, 2012; Else-Quest & Hamilton, 2018). For example, researchers are advised to select gender equality indicators related to education (e.g., enrollment ratios, literacy rates, women's share of higher positions, or women's share of research positions) when examining gender gaps in educational outcomes rather than composite gender equality indicators (such as the GGGI). This will result in more meaningful results and ultimately a higher validity of the body of knowledge on the impact of gender equality in the literature.

4.3 Research Question III: Which Functional Relation Exists Between Students' Academic Achievement and Corresponding Academic Self-Concepts?

Study II examined the functional relation between academic achievement and corresponding academic self-concepts across (a) different domains (i.e., mathematics and verbal), (b) different age groups (i.e., elementary and secondary school students), and (c)

13 different countries. To this end, we drew on representative individual student data from eight assessment cycles of three major educational large-scale studies (i.e., TIMSS, PIRLS, PISA; $N = 470{,}804$) and applied quadratic and interrupted regression analyses. We combined the results from each country and cycle in an integrative data analysis to examine their generalizability and robustness (see Curran & Hussong, 2009; Hofer & Piccinin, 2009).

4.3.1 Results From Study II

The findings provided strong evidence of nonlinear relations between achievement and self-concepts for students in secondary school in mathematics and the verbal domain. For secondary school students, the results from the quadratic regression analyses and interrupted regression analyses implied that the increase in students' self-concept was weaker for lower achieving students than for higher achieving students in both domains. For 15-year-old students, the interrupted regression analysis even revealed that mathematics achievement was not significantly related to self-concept of mathematics for lower achieving students, but the relation was significantly positive for higher achieving students.

Nonlinear effects were also present in elementary school students, but the pattern of results was rather heterogeneous across countries and applied methods. On average, the findings from the quadratic regression analyses indicated that—similar to the results for older students—the increase in students' mathematics self-concept in Grade 4 was lower for lower achieving students than it was for higher achieving students. However, the interrupted regression analyses showed that the relation between mathematics achievement and mathematics self-concept differed not much for higher and lower achieving students. A different results pattern emerged for the verbal domain, where the quadratic regression

analyses indicated that reading achievement and reading self-concept were to some extent more strongly related for lower achieving students than for higher achieving students in Grade 4. The interrupted regression analyses confirmed this result. Thus, the findings of the study were mixed regarding age differences in nonlinear relations between achievement and self-concepts in mathematics and the verbal domain (see also the Discussion section in Study II).

To conclude, Study II shows that a linear regression model could not fully capture the relations between achievement and the corresponding self-concepts in mathematics and the verbal domain for a substantial proportion of the student body in secondary school. As indicated by the interrupted regression models, this "substantial proportion" ranged from, on average, 12% (TIMSS, Grade 8) to 22% (PISA, verbal domain) of the students. Given the heterogeneity in the results as observed for elementary school students, we recommend that nonlinear models not only be specified for secondary school students but also for elementary school students.

4.3.2 From Tools to Theories: How Do Statistical Models Influence Our Scientific Knowledge Gain?

In his tools-to-theories heuristic, Gigerenzer (1991) argues that scientific tools (i.e., methods and instruments) inspire researchers to new theoretical metaphors. For instance, in the 19th century, Faraday's instruments for recording electric currents influenced the understanding of electrophysiological processes, which is reflected in concepts such as "muscle current" and "nerve current" (Lenoir, 1986). According to Gigerenzer (1991), an example for the influence of scientific methods on theory development is the institutionalization of inferential statistics in the 1960s. The introduction of inferential statistics into cognitive psychology led to a reinterpretation of many cognitive processes

suggesting that the mind is an "intuitive statistician." For example, Kelley (1967) assumed in his causal attribution theory that the mind attributes a cause to an effect by conducting an ANOVA and testing null hypotheses (Gigerenzer, 1991). However, Gigerenzer (1991) also stresses that the fact that our scientific instruments and statistical models shape scientific theory development should also make us aware of the limitations of current theories and research programs and for limitations in the further development of alternatives and new possibilities. Gigerenzer's conclusions can also be applied to other research areas, such as the research on academic self-concepts. Similar to the "ANOVA mind," there might be a "linear regression mind" in researchers studying academic self-concepts and its formation. As demonstrated in the Introduction section in Study II, the implicit assumption of a linear relation between academic achievement and corresponding self-concepts is prevalent in educational psychology. However, the results in Study II show that assuming a nonlinear relation between achievement and corresponding self-concepts is highly plausible and might better capture the relation between those two constructs, at least for secondary school students. An adherence to the "linear regression mind" might hamper theory development in the long run.

4.3.3 Improving the Testability of Theories by Specifying Functional Relations

According to Popper (2002), the higher the probability with which a theory can be falsified, the higher its scientific quality and value. Popper calls this probability of falsification the *informative content* of a theory. Thus, the higher the informative content of a theory, the higher its scientific quality and value, as it offers more possibilities for testability and falsification. Indeed, Meehl (1967) criticized the lack of informative content of psychological theories as an obstacle to progress in psychological research.

The form of the functional relation between academic achievement and corresponding self-concepts is usually not specified in theories that involve these constructs such as the SEVT (e.g. Eccles & Wigfield, 2020), or the I/E Model (Marsh, 1986), the REM Model (Marsh & Martin, 2011), the Dimensional Comparison Theory (Möller & Marsh, 2013), or the BFLPE Model (Marsh, 1987). However, specifying assumptions on the functional relation between academic achievement and corresponding self-concepts would increase the informative content of these theories (or models) and would consequently enhance their scientific quality.

4.3.4 Measuring the Influence of Response Styles on Academic Self-Concepts by Using Vignette Formats

As previously defined, academic self-concepts represent a person's mental representations of his or her own abilities in academic domains (Marsh & Craven, 1997). As a result, they are only accessible through introspection and are assessed by self-reports. The assessment of academic self-concepts relies on Likert items, which present students with statements to which they respond to by ticking a box on a scale (e.g., strongly disagree, disagree, agree, strongly agree). These items, while convenient, efficient, and highly predictive of key life outcomes (Duckworth & Yeager, 2015), have a variety of response biases that can impact the validity of scores obtained (Shadish et al., 2002). As briefly discussed in Study II, response styles bias how individuals respond to Likert items and can consequently influence the functional relations between achievement and corresponding self-concepts. Response styles refer to the tendency to agree with most items, regardless of their content (acquiescence), to disagree with most items, regardless of their content (disacquiescence), to use the endpoints of a scale (extreme response style), and to carelessly of randomly answer items (noncontingent responding; Buckley, 2009; Kyllonen, 2016). Buckley (2009)

showed that adjusting for students' response biases tended to result in 2 out of 3 studied countries in weak negative nonlinear relations between science achievement and a self-constructed global science attitude scale in PISA 2006.

A promising way to control for response style effects is to rescale items based on an anchoring vignettes approach (e.g., Bolt et al., 2014; He et al., 2017; King et al., 2003; King & Wand, 2007). In this approach, individuals are asked to rate so-called vignettes that, for example, describe low, medium, and high levels of teacher support behavior on a categorical rating scale (OECD, 2014; Table 3).

Table 3. *Anchoring Vignettes Based on Teacher Support Behaviors in PISA 2012*

Level of TS	Description
Low level	Ms. <name> sets mathematics homework once a week. She never gets the answers back to students before examinations.
Medium level	Mr. <name> sets mathematics homework once a week. He always gets the answers back to students before examinations.
High level	Ms. <name> sets mathematics homework every other day. She always gets the answers back to students before examinations.

Note. TS = Teacher support. 4-point rating scale (*strongly disagree* to *strongly agree*). OECD (2014), p. 52.

It is supposed that systematic differences in individuals' ratings of the same vignette mainly reflect differences in response styles, whereas individuals' ratings of the target item are a combination of response style distortion and the true trait level. Thus, to obtain a response-style-free estimate of the actual level of the target trait, individuals' ratings of the vignettes are in the next step used as personal standards by which responses to the target items are rescaled (He et al., 2017). In previous studies, the vignette format has been successfully applied to reduce the influence of response styles on variables measured with Likert items (e.g., He et al., 2017; Mõttus et al., 2012; Primi et al., 2016). The vignette format was also applied in PISA 2012 on teacher support and classroom management

scales (OECD, 2014). Thus, this technique could be used to examine the robustness of part of the findings from Study II against the influence of response styles (i.e., the nonlinear relations between achievement and self-concept in mathematics in PISA 2012). Some researchers propose that for reasons of efficiency the same set of anchoring vignettes can be applied to different constructs of the same response format because the response styles to different constructs remain the same (e.g., Kyllonen & Bertling, 2014). However, other researchers assume that individuals' personal standards depend on the specific target construct and would therefore advise against this proposal and recommend to use vignettes closely related to the construct of interest (e.g., He et al., 2017; Vonkova et al., 2017). To conclude, using the teacher support and classroom management vignettes to rescale students' self-concept responses could be a first, but maybe not yet perfect, attempt to test the robustness of the nonlinear effects against response style bias.

4.4 Strengths, Limitations, and Directions for Future Research

Study-specific strengths and limitations are discussed in the Discussion section of each study (see Chapters 2 and 3). In the following, strengths and limitations of the doctoral thesis as a whole will be highlighted and, subsequently, directions for future research will be discussed.

4.4.1 Strengths and Limitations

The present doctoral thesis comprises both strengths and limitations.

A first strength is that I was able to use international large-scale assessment that provide high quality data. The data quality is characterized by the fact that TIMSS, PIRLS, and PISA provide student data from defined populations (i.e., students in Grade 4 and 8 in TIMSS, students in Grade 4 in PIRLS, and 15-year-olds in PISA) that were

representatively sampled. Furthermore, measuring instruments in these assessments are rigorously tested in field studies to ensure high psychometric quality. Most importantly, the exclusive use of data from international large-scale assessments considerably reduced the heterogeneity of effect sizes in both studies due to greater methodological homogeneity (e.g., applying the same measures across cycles). Thus, using data from international large-scale assessments in Study I and II provided the strongest data basis to answer the three research questions.

Second, following the principles of critical multiplicism (Shadish, 1993), the functional relations between achievement and self-concepts in Study II were examined with different analytical methods, in different populations, countries, and domains, in order to minimize bias.

Third, and as one focus of this thesis, state-of-the-arts methods of data analyses were used in both studies to examine the interplay of achievement and achievement motivation. For example, research synthesis methods, such as individual participant data meta-analysis and integrative data analysis, as well as multiple imputation techniques to impute missing data (Study I: missing data on moderator variables, Study II: missing data on self-concepts) were used. Especially the use of research synthesis methods that synthesized effect sizes across multiple countries and cycles of PISA, TIMSS, and PIRLS yielded strong (i.e., robust and generalizable) evidence for gender differences in top-performing math students, moderating influences of specific gender equality indicators, and nonlinear relations between achievement and corresponding self-concepts.

Fourth, and specifically related to Study I, many facets of gender differences in achievement, achievement profiles, and motivation were systematically investigated together for the first time in a representative sample of top-performing math students. These can be considered as antecedents of gender differences in educational and

occupational STEM choices. Thus, Study I provided a comprehensive contribution to the field of gender and STEM.

Taken together, both Studies I and II as well as the doctoral thesis in general provide strong empirical evidence on the interplay of achievement and achievement motivation.

Despite its strengths, the present doctoral thesis is not without general limitations, which I will discuss in the following. One limitation might be that large-scale assessments cannot explain why outcomes in individual countries are as they are because their design warrants no causal conclusions. For example, they cannot provide any concrete guidelines on how to ensure greater gender equality in education for specific countries (Study I), or inform us why the functional relations between achievement and self-concept varied across countries (Study II). Additional information needs to be collected in these cases (e.g., interviews with education experts or politicians about characteristics of national education systems). Furthermore, despite the many advantages that come with using data from international large-scale assessments (see above), another limitation might be that the exclusive use of data from international large-scale assessments entails the risk that specific characteristics of large-scale assessments could systematically influence the results of the study; these characteristics might include low-stakes testing (i.e., assessments have no influence on students' course grades), which might influence students' motivation in these assessments; different conceptual frameworks may be involved (e.g., focus on literacy in PISA vs. on the curriculum in TIMSS and PIRLS might yield different relations between achievement and motivational variables); and the introduction of computer-based assessment in PISA 2015 potentially induce mode effects (i.e., items are systematically easier or harder when delivered on computer as compared to the paper-based assessment) that might differ between male and female students (Jerrim et al., 2018).

Nonetheless, the aforementioned advantages of large-scale assessments outweigh their disadvantages by far.

4.4.2 Directions for Future Research

In the next sections, I will outline four research avenues for rendering analyses in the area of the present doctoral thesis more multiverse, robust, and specific.

4.4.2.1 Context-Sensitive Profiles in Achievement and Achievement Motivation

As argued before, scientific findings are shaped by the statistical models we apply in our research. Something similar could be observed in research that uses achievement and achievement motivation profiles. In the Introduction of this doctoral thesis, I presented several approaches on how profiles in achievement and achievement motivation can be operationalized (Section 1.5.1). These approaches differed in their focus (variable-centered vs. person-centered), in their breadth (profiles of two or more constructs), and how the criteria for the profile formation were determined (content-related or data driven). The way profiles are operationalized likely affects the results of a study.

As intraindividual hierarchies in achievement and achievement motivation play an important role in the SEVT for predicting individuals' future achievement and achievement-related choices (Eccles, 1994; Wigfield & Eccles, 2020), a systematic overview of profiles in achievement and achievement motivation would be highly desirable. This overview could cover definitions of profiles and develop a taxonomy that defines in which contexts for which research purposes which profiles should be used.

4.4.2.2 Robust Analyses of Nation-Level Gender Equality

Researchers interested in examining the influence of gender equality on gender differences in educational outcomes are faced with a range of analytical decisions that will likely affect the results of their study. These include, for example, from which period of time the gender equality indicators should be included, how the mapping between nation-level indicators and countries or economic regions is done, whether or not different data sets of the same indicator are combined to increase the available information on countries, whether or not several data sets from international large-scale assessments are combined to increase the sample of countries, and whether or not missing data on the indicators are imputed. In the following, I will briefly describe typical analytic decisions in this field.

There are at least two possible approaches to the question of the period of time for which the indicators should be included. One approach is to match the indicator with the year in which the target variables were collected (e.g., Penner, 2008; Stoet & Geary, 2018) or to choose the year closest in time to the collection of the target variables if data from the same year were not available (e.g., Else-Quest et al., 2010; Guiso et al., 2008; Machin & Pekkarinen, 2008; Reilly, 2012; Riegle-Crumb, 2005; Stoet & Geary, 2015). For example, Riegle-Crumb (2005) drew on data from TIMSS 1995 and chose gender equality indicators from "the early 1990s" (p. 231) to be "representative of the extent of stratification that existed at approximately the same time that the TIMSS survey were being administered to students" (pp. 241–242). Another approach is to use indicators from previous years (e.g., Baker & Jones, 1993). For example, Baker and Jones (1993) drew on data from the Second International Mathematics Study (SIMS) from 1982 and gender equality indicators from the years 1970 to 1975 because "these indicators show the stratification of opportunity

before the collection of the SIMS data or the adult opportunity structure evident to students in the SIMS study" (p. 95; Baker & Jones, 1993).

Another analytical decision is the mapping between nation-level indicators and the countries or economic regions that participated in the large-scale assessments. Whereas domain-specific gender equality indicators are usually only available at the nation level, large-scale assessments (such as PISA) also collect data from students in specific economic regions (e.g., Beijing, Shanghai, Jiangsu, and Guangdong in China, PISA 2015). It is up to the researchers to decide whether to exclude the economic regions from the analyses (e.g., Stoet & Geary, 2020b), whether the characteristics of the economic regions and associated nations are similar enough to replace the missing regional values with the national value (e.g., Study I), or to treat the data as missing and apply a missing data imputation technique (e.g., multiple imputation).

Furthermore, different organizations provide different data on the same gender equality indicators (different ranges of countries and different time periods). For example, women's share of research positions is both documented by the UNESCO and the OECD. Thus, researchers may choose to combine these data in different ways. They can (1) treat data sets equally by aggregating all available information (e.g., Stoet & Geary, 2015; Study I), (2) relying on information from one data set and filling only the gaps with data from the other data set, or (3) just use data from one of the data sets (e.g., Else-Quest et al., 2010).

In addition, some researchers analyzed single cycles from international large-scale assessments to answer their research questions (e.g., Else-Quest et al., 2010; Guiso et al., 2008; Machin & Pekkarinen, 2008) or treated several cycles separately (Stoet & Geary, 2013, 2015, 2018), whereas other researchers (Study I) aggregated several cycles from international large-scale assessments. The aggregation increases the number of countries in

the sample and thus may enhance the statistical power and precision to estimate the relation between gender differences and gender equality indicators.

Finally, results might vary depending on whether missing data on gender equality indicators were imputed (e.g., Study I) or not (e.g., Else-Quest et al., 2010; Guiso et al., 2008; Machin & Pekkarinen, 2008; Penner, 2008; Stoet & Geary, 2013, 2015, 2018). Imputing missing data may also increase the statistical power and precision as well as mitigate bias when estimating the relation between gender differences and gender equality indicators.

Given the different analytical approaches that can be found in the literature on the relation between gender equality and gender differences in educational outcomes, it would be interesting to conduct a specification curve analysis (Simonsohn et al., 2015) or multiverse analysis (Steegen et al., 2016). Both specification curve analysis and multiverse analysis assess the robustness of findings by performing and combining all theoretically justified, statistically valid, and non-redundant analyses across all alternative data sets.

4.4.2.3 From Nation-Level Gender Equality to Specific Regional-Level Gender Equality

Study I investigated the moderating effects of several domain-specific gender equality indicators on the share of female students in the top 5% in mathematics and gender differences within this group of students. It was found that tertiary enrollment ratios predicted the proportion of female students in the top 5% in mathematics, and tertiary enrollment ratios and women's share of research positions in a country predicted mathematically top-performing female and male students' achievement profile scores. However, these patterns tell us little about the causal relationships between nation-level gender equality indicators and gender differences in this group of students. A longitudinal

research design that tracks gender equality indicators and gender differences in (top-performing math) students' achievement (and possibly their achievement profiles and achievement motivation) over time might better approximate the estimation of causal relationships. Since the available PISA data basis is continuously expanded by one cycle every three years (e.g., students' mathematics achievement can currently be linked across a period of 15 years [PISA 2003–2018]), this would be an interesting research endeavor.

In addition, most research on the influence of gender equality on gender differences in educational outcomes has been conducted at the international level. However, gender equality also differs regionally (e.g., Bundesministerium für Familie, Senioren, Frauen und Jugend [BMFSFJ], 2017). How children and adolescents experience gender equality in their more immediate environment is likely to have a greater impact on their development than the average gender equality at national level. To advance gender equality within a country, it would be useful to examine the relation between gender equality and gender differences in educational outcomes of female and male students also at the regional level. The results of such a study could be used to support evidence-based policy making. For example for Germany, a multilevel (longitudinal) study could be conducted by matching data on students' educational outcomes from national assessment studies (e.g., the IQB's National Assessment Study and IQB Trend in Student Achievement) and gender equality indicators (such as women's share of higher positions, women's share of elected officials in municipal representations, part-time employment rate for women, employment rate of mothers with young children, father's share of parental benefit) available for the years 2008, 2011, and 2015 (BMFSFJ, 2010, 2013, 2017) at the state level.

4.4.2.4 Gender as Nonbinary Variable in Psychological and Educational Research

In the data that I used in the present doctoral thesis, gender was conceptualized as binary. That is, answering the student questionnaire in TIMSS, PIRLS, and PISA, students could categorize themselves into just two categories: female or male. This approach is based on what is referred to as the *gender binary*. Psychologists have justified this dichotomization by the fact that men and women could be characterized by separate sets of brain features, hormones, psychological characteristics, and gender identities. This practice has been considered as natural and inevitable for decades (Hyde et al., 2019). However, different forces challenge psychology's assumption of the gender binary, ranging from the transgender activist movement (e.g., Stryker, 2008), the intersex activist movement (e.g., Reis, 2007) to recent research findings. For example, evidence from neuroscience and behavioral endocrinology refutes the gender dimorphism of the human brain (e.g., Joel et al., 2015) and the hormonal systems (e.g., Gillies & McArthur, 2010); findings from psychology emphasize the similarities between women and men (e.g., Zell et al., 2015); developmental research stresses the social-cognitive mechanisms with which we learn that gender is a culturally meaningful category as children (e.g., Bigler, 2013); and psychological research documents transgender and nonbinary individuals' identities and experiences (e.g., Tate et al., 2014; for reviews, see Hyde et al., 2019; Schellenberg & Kaiser, 2018). As a result, more and more countries adopt laws to accommodate nonbinary gender identities (Schmidt & Fox, 2018).

To study this rich complexity of gender in adolescents and adults, researchers need to assess individuals' gender in nonbinary ways. In the following, two methods are presented to do so. One method is to provide individuals with several options to choose from when assessing their gender. These options could include "female," "male,"

"transgender female," "transgender male," "genderqueer," and "other (specify)." Alternatively, researchers can ask individuals to self-identify by using open-ended questions (e.g., "What is your gender?" or "How do you currently identify?"). Furthermore, researchers might assess birth-assigned and self-assigned gender identities separately (Hyde et al., 2019; Lindqvist et al., 2020).

A second method to assess individuals' gender in a nonbinary way is to treat gender as a multidimensional, continuous construct. One means to do this is to use inventories, such as the Bem Sex-Role Inventory (BSRI; Bem, 1974), the Personal Attributes Questionnaire (PAQ; Spence & Helmreich, 1978), or the Multi-Gender Identity Questionnaire (Multi-GIQ; Joel et al., 2014). The BSRI assesses femininity and masculinity as two independent dimensions on which an individual can score low or high on both dimensions or high on one dimension and low on the other (but see also Hoffman & Borders, 2001). The PAQ measures the degree to which a person can be classified according to masculine (agentic) or feminine (communal) adjectives and consists of three scales: the instrumentality scale (masculinity), the expressivity scale (femininity), and the androgyny scale (masculinity–femininity; Spence & Helmreich, 1978). The Multi-Gender Identity Questionnaire (Multi-GIQ; Joel et al., 2014) assesses individuals' self-identification with femininity and masculinity on levels related to gender identity, gender expression, legal gender, and bodily aspects (for reviews, see Lindqvist et al., 2020; Wood & Eagly, 2015). The most comprehensive approach to measuring gender is proposed by Schellenberg and Kaiser (2018) who suggest six gender dimensions (gender expression, gender identification, gender attitudes, gender traits, [recalled] gender socialization, and gender role behavior).

Future research on gender differences in adolescents should strive for a more nuanced view on gender. Assessing gender in a nonbinary way and incorporating

dimensional gender measures, not only contributes to a more inclusive society, but also expands our understanding of the underlying mechanisms that produce or moderate gender differences (e.g., hormone levels, gender role expectations, etc.). There are already some examples of studies that take a more nuanced approach. For example, a study by Reilly et al. (2016) investigated whether gender role identity mediates the relationship between gender and gender-typed cognitive abilities. However, to date, research rarely assesses individuals' gender in a nonbinary way. Incorporating nonbinary measures for gender into large-scale studies, such as TIMSS or PISA, could make an important contribution, since large-scale studies often have much larger sample sizes than individual studies and are usually drawn on a representative basis.

4.5 Implications for (Educational) Policy and Practice

In the following sections, I will discuss selected practical implications that arise from the results of the present doctoral thesis. These implications address (1) how to increase women's representation in STEM, and (2) what nonlinear relations between achievement and corresponding self-concepts imply for achievement-based interventions.

4.5.1 How to Increase Women's Representation in STEM?

The results of Study I showed that very many mathematically able female students do exist. Although the female-to-male ratio in the top 5% in mathematics was not perfectly balanced (female-to-male ratio 1:1.50), the study supports the conclusion that there is a considerable number of women in the STEM pool in almost all countries. Thus, what can be done to increase women's representation in STEM careers, especially in the fields of engineering, physics, and computer sciences that are characterized by the most extreme gender disparities?

From a historical perspective, programming, for example, has not always been considered a male domain. In the early 1940s, the world's first computer programmers who worked on the ENIAC—the first large, general-purpose electronic computer—were women (Morell, 1996). In 1967, in an article titled "The Computer Girls" even the women fashion and entertainment magazine *Cosmopolitan* advertised programming as an ideal job for women and described the profession as offering better job opportunities for women than many other professional careers. However, already since the 1950s, computer programming was increasingly "made masculine" as the field was beginning to acquire new status. One reason why women were driven out of this field was, for example, the development of the notion that programmers typically show no interest in people. Consequently, "disinterest in people" had been taken into account as a criterion in personality tests for personnel selection by the majority of companies in the mid-1960s, which resulted in an increasing male dominance of the field (Ensmenger, 2010).

As noted before, the SEVT assumes that women are less likely to enter STEM fields than men because they have lower expectancies for success and lower subjective task values with regard to STEM as compared to other fields (e.g., the humanities; Eccles, 1994). According to Diekman et al.'s (2011) goal congruity model, there are less women in STEM because STEM careers are not perceived as the best choice for fulfilling the communal goals that are especially valued by women (e.g., working with or helping others). This perceived mismatch between women's communal goals and their belief that STEM careers do not embody such values may ultimately result in female students' particular disinterest in STEM fields (Diekman et al., 2011). Based on the SEVT (Eccles & Wigfield, 2020; Wigfield & Eccles, 2020) and the goal congruity model (Diekman et al., 2011), four strategies will be presented that might help to (re)recruit women into STEM: (1) change STEM culture by changing STEM stereotypes, (2) enhance students'

expectancies for success and values regarding STEM, (3) align STEM activities with students' values, and (4) reduce the perceived importance of brilliance in STEM (see also Diekman et al., 2019).

4.5.1.1 Change STEM Culture by Changing STEM Stereotypes

Especially in the fields of computer science, engineering, and physics, the stereotypical perceptions are that the people working in these fields are male, socially awkward, and focused on technology (Cheryan et al., 2015, 2017). Thus, women who enter these contexts may be less likely to be considered as belonging to STEM, which is also reflected in their experiences (Diekman et al., 2019). These stereotypes can be challenged by knowing people in the field who do not confirm these stereotypes (e.g., Cheryan et al., 2013). For example, available female role models and mentors can have positive effects on female students' STEM engagement (Downing et al., 2005). Female role models can be STEM practitioners whose work is highlighted in courses or who visit the school or the university, or teachers or faculty within schools or STEM departments.

4.5.1.2 Enhance Students' Expectancies for Success and Values Regarding STEM

Drawing on the SEVT, a plethora of studies have shown that utility value interventions and also cost interventions can increase students' interest, performance, and attainment in STEM at high school and university level in the short and long term (e.g., Gaspard et al., 2015; Hecht et al., 2019; Hulleman & Harackiewicz, 2009; Rosenzweig et al., 2020; Rozek et al., 2015). Furthermore, it has been shown that utility value interventions can also positively influence other motivational beliefs and values (Hecht et al., 2019). These interventions mostly consist of students writing an essay on the relevance of specific STEM fields to their personal lives (utility value intervention) or on how they could

perceive the challenges of their physics course as less psychologically costly (cost intervention; Gaspard et al., 2015; Hecht et al., 2019; Hulleman & Harackiewicz, 2009; Rosenzweig et al., 2020). Hence, these findings suggest that it might be useful to integrate essays on the relevance of STEM for students' own lives into (e.g., the STEM) curricula at school and university to foster female students', but also male students', STEM participation. Importantly, an intervention that targeted parents' utility value beliefs about math and science led to an increased course-taking by high-achieving female students (Rozek et al., 2015). Thus, informing high-achieving adolescent girls' parents about the usefulness of mathematics and science for adolescents seems also to be an effective tool for increasing female students' STEM engagement.

4.5.1.3 Align STEM Activities With Students' Values

From the communal goal congruity perspective, another way to make STEM fields more attractive to girls and women could be to frame STEM fields as affording a wider range of goals, especially communal goals (Diekman et al., 2019). For example, research shows that information or experiences that convey how STEM can fulfill communal goals can have beneficial effects on STEM motivation and positivity for individuals who value communal goals, irrespective of their gender. Already brief exposures to scientist exemplars engaged in communal work (e.g., collaborating with colleagues) fostered beliefs that STEM could meet communal goals and positive attitudes toward pursuing STEM careers in women (Clark et al., 2016; Diekman et al., 2011; Diekman & Fuesting, 2018). Furthermore, a hypothetical lab experience with communal opportunities (i.e., opportunities for face-to-face connection and mentoring) cued anticipated belonging and interest in joining the lab in undergraduate STEM majors (Belanger et al., 2020). Thus, designing STEM courses at school or university that provide communal experiences by integrating collaborative work

among students, by conducting projects that help others (e.g., developing online systems for medical or psychological counseling), or by inviting or visiting scientists who highlight in which ways their work is communal would be a means to increase female students' engagement in STEM. These experiences are important because they dispel beliefs (with a kernel of truth) that STEM fields are solitary and competitive (Diekman et al., 2019).

4.5.1.4 Reduce the Perceived Importance of Brilliance in STEM

From the early decades of computing, programming, for example, was seen as a "black art" that requires an innate aptitude (Ensmenger, 2010). Recent research on field-specific ability beliefs suggests that women are underrepresented in fields whose members believe that raw, innate talent is the main requirement for success, because women are stereotyped as not possessing such talent (Bian et al., 2018; Leslie et al., 2015; Meyer et al., 2015; Storage et al., 2016, 2020). There is evidence that children at elementary school level already associate brilliance more strongly with boys than with girls (Bian et al., 2017). Importantly, messages about the significance of brilliance to success in a field, rather than dedication, reduced women's, but not men's, interest (Bian et al., 2018).

To reduce the perceived importance of brilliance in STEM, a first step might be to minimize the talk of genius or brilliance with students to make a field more welcoming. Furthermore, teachers and faculty could encourage a growth perspective in students, as opposed to a fixed-trait mindset (Cimpian & Leslie, 2017; see Dweck, 1999, 2006). Research shows that describing a novel STEM field as focused on effort increased women's feelings of belonging and future motivation (Smith et al., 2013). Furthermore, perceiving others in a calculus class at university level to have a growth mindset about math ability allowed female students to maintain a high sense of belonging in math and the intention to pursue math in the future. On the contrary, a fixed mindset about math ability

reduced female students', but not male students', sense of belonging and future math intentions, especially if they also perceived others to endorse negative stereotypes about women's math abilities (Good et al., 2012). A growth mindset could be fostered, for example, by more advanced students, teachers, faculty, visiting scientists, or STEM practitioners sharing personal experiences. This may be especially effective when they struggled in STEM and share how they overcame these obstacles emphasizing that such struggles are inherent to challenging work and not to a lack of talent (Aronson et al., 2002; Diekman et al., 2019).

To conclude, the most promising way to increase women's representation in STEM careers seems to be to change the STEM culture (Cheryan et al., 2017; Cimipian et al., 2020). Broadening the representation of the people who work in STEM, broadening the work itself, establishing a growth mindset, and expanding the environments in which STEM professionals work may change the stereotypes associated with STEM fields and might further increase girls' and women's sense of belonging and interest in STEM (Cheryan et al., 2015, 2017; Diekman et al., 2019).

4.5.2 Implications for Achievement-Related Interventions

The findings in Study II indicated that the relation between academic achievement and corresponding self-concepts was not linear across the whole achievement distribution. For example, for 15-year-old students, mathematics achievement and mathematics self-concepts were not related in lower achieving students (i.e., on average for 16% of 15-year-olds at the lower end of the achievement distribution). However, for higher achieving students, mathematics achievement and mathematics self-concept were positively related. This nonlinear relation could potentially have implications for interventions aimed at improving the skills of children and adolescents at risk.

A typical problem of interventions is that so-called fade-out effects occur (Abenavoli, 2019; Bailey et al., 2017; Protzko, 2015). This means that the effect of the intervention disappears after some time. An important question for intervention researchers is therefore how sustainable effects of interventions can be achieved. Based on the results in Study II, it could be speculated that intervention effects are not sustained for lower achieving students, because their self-assessment might be disconnected from their achievement (e.g., due to self-protection strategies). If students' achievement is not linked to their self-assessment, this can have negative consequences, for example, for successful self-regulated learning processes, which in turn might have negative effects on their achievement in the long term (e.g., Sticca et al., 2017). Consequently, the positive intervention effects might not be sustainable. This is problematic because it means that achievement-related interventions may not work for those students who need them most. To reconnect lower achieving students' achievement and corresponding self-concepts, one approach might be to also target their self-concepts in achievement-related interventions. This could be done, for example, by incorporating appropriate praise and/or feedback strategies into the intervention. These strategies should be contingent upon performance that is attributional in nature (i.e., students should learn to attribute success internally and failure externally) and goal-relevant (O'Mara et al., 2006).

4.6 References

Abelson, R. P. (1985). A variance explanation paradox: when a little is a lot. *Psychological Bulletin*, *97*(1), 129–133. https://doi.org/10.1037/0033-2909.97.1.129

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin*, *145*(12), 1103–1127. http://dx.doi.org/10.1037/bul0000212

Achter, J. A., Lubinski, D., & Benbow, C. P. (1996). Multipotentiality among the intellectually gifted: "It was never there and already it's vanishing." *Journal of Counseling Psychology, 43*(1), 65–76. https://doi.org/10.1037/0022-0167.43.1.65

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*(2), 113–125. https://doi.org/10.1006/jesp.2001.1491

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48*(4), 217–228. https://doi.org/10.3102/0013189X19848729

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66*(2), 91–103. https://www.jstor.org/stable/2112795

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education, 4*(1). https://doi.org/10.1186/s40536-015-0015

Belanger, A. L., Joshi, M. P., Fuesting, M. A., Weisgram, E. S., Claypool, H. M., & Diekman, A. B. (2020). Putting belonging in context: Communal affordances signal belonging in STEM. *Personality and Social Psychology Bulletin,* 1–19, https://doi.org/10.1177/0146167219897181

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155–162. https://doi.org/10.1037/h0036215

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391. https://doi.org/10.1126/science.aah6524

Bian, L., Leslie, S. J., & Cimpian, A. (2018). Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist, 73*(9), 1139–1153. http://dx.doi.org/10.1037/amp0000427

Bigler, R. S. (2013). Understanding and reducing social stereotyping and prejudice among children. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 327–331). Oxford University Press. https://dx.doi.org/10.1093/acprof:oso/9780199890712.003.0060

BMFSFJ (2010). *Atlas zur Gleichstellung von Frauen und Männern in Deutschland [Gender Equality Atlas for Germany].* https://www.bmfsfj.de/blob/93282/54bc2ae0e85ce1d1a6484d948d2233bc/atlas-gleichstellung-deutschland-data.pdf

BMFSFJ (2013). *2. Atlas zur Gleichstellung von Frauen und Männern in Deutschland [2nd Gender Equality Atlas for Germany].*

https://www.bmfsfj.de/blob/93150/370b717c3ddc5f7235b56e0a3d987bf1/2--atlas-zur-gleichstellung-in-deutschland-data.pdf

BMFSFJ (2017). *3. Atlas zur Gleichstellung von Frauen und Männern in Deutschland [3rd Gender Equality Atlas for Germany].* https://www.bmfsfj.de/blob/114006/738fd7b84c664e8747c8719a163aa7d9/3--atlas-zur-gleichstellung-von-frauen-und-maennern-in-deutschland-deutsch-data.pdf

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528–541. https://doi.org/10.1037/met0000016

Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Cheryan, S., Drury, B. J., & Vichayapai, M. (2013). Enduring influence of stereotypical computer science role models on women's academic aspirations. *Psychology of Women Quarterly, 37*(1), 72–79. https://doi.org/10.1177/0361684312459328

Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology, 6*, 49. https://doi.org/10.3389/fpsyg.2015.00049

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143*(1), 1–35. https://dx.doi.org/10.1037/bul0000052

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science, 368*(6497), 1317–1319. https://doi.org/10.1126/science.aba7377

Cimpian, A., & Leslie, S. J. (2017). The brilliance trap. *Scientific American, 317*(3), 60–65.

Clark, E. K., Fuesting, M. A., & Diekman, A. B. (2016). Enhancing interest in science: Exemplars as cues to communal affordances of science. *Journal of Applied Social Psychology, 46*(11), 641–654. https://doi.org/10.1111/jasp.12392

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Academic Press.

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

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist, 64*(3), 170–180. https://doi.org/10.1037/a0014564

Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science, 12*(3), 258–261. https://doi.org/10.1111/1467-9280.00347

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond.* Routledge.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Dabney, K. P., Tai, R. H., Almarode, J. T., Miller-Friedmann, J. L., Sonnert, G., Sadler, P. M., & Hazari, Z. (2012). Out-of-school time science activities and their association

with career interest in STEM. *International Journal of Science Education, Part B, 2*(1), 63–79. https://doi.org/10.1080/21548455.2011.629455

Diekman, A. B., Brown, E., Johnston, A., & Clark, E. (2010). Seeking congruity between goals and roles: A new look at why women opt out of STEM careers. *Psychological Science, 21*(8), 1051–1057. https://doi.org/10.1177/0956797610377342

Diekman, A. B., Clark, E. K., & Belanger, A. L. (2019). Finding common ground: synthesizing divergent theoretical views to promote women's STEM pursuits. *Social Issues and Policy Review, 13*(1), 182–210. https://doi.org/10.1111/sipr.12052

Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., & Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to STEM careers: Evidence for a goal congruity perspective. *Journal of Personality and Social Psychology, 101*(5), 902–918. https://doi.org/10.1037/a0025199

Diekman, A. B., & Fuesting, M. A. (2018). Choice, context, and constraint: When and why do women disengage from STEM? In C. B. Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA handbook of the psychology of women: Perspectives on women's private and public lives* (pp. 475–495). American Psychological Association. https://doi.org/10.1037/0000060-026

Downing, R. A., Crosby, F. J., & Blake-Beard, S. (2005). The perceived importance of developmental relationships on women undergraduates' pursuit of science. *Psychology of Women Quarterly, 29*(4), 419–426. https://doi.org/10.1111/j.1471-6402.2005.00242.x

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development.* Psychology Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* Random House.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Erlbaum.

Eagly, A. H. (2013). The science and politics of comparing women and men: A reconsideration. In M. K. Ryan & N. R. Branscombe, *The SAGE handbook of gender and psychology* (pp. 11-28). SAGE Publications Inc. https://doi.org/10.4135/9781446269930.n2

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist 75*(3), 301–315. https://doi.org/10.1037/amp0000494

Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of Women Quarterly, 18*(4), 585–609. https://doi.org/10.1111/j.1471-6402.1994.tb01049.x

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61*, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Else-Quest, N. M., & Grabe, S. (2012). The political is personal: Measurement and application of nation-level indicators of gender equity in psychological research. *Psychology of Women Quarterly, 36*(2), 131–144. https://doi.org/10.1177/0361684312441592

Else-Quest, N. M., & Hamilton, V. (2018). Measurement and analysis of nation-level gender equity in the psychology of women. In C. B. Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA handbook of the psychology of women: Perspectives on women's private and public lives* (p. 545–563). American Psychological Association. https://doi.org/10.1037/0000060-029

Else-Quest, N. M., Hyde, J. S., Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Ensmenger, N. (2010). Making programming masculine. In T. J. Misa (Ed.), *Gender codes: Why women are leaving computing* (pp. 141–130). Wiley/IEEE Computer Society. https://dx.doi.org/10.1002/9780470619926.ch6

Gaspard, H., Dicke, A.-L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology, 51*(9), 1226–1240. https://doi.org/10.1037/dev0000028

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*(2), 254–267. https://doi.org/10.1080/14792779143000033

Gillies, G. E., & McArthur, S. (2010). Estrogen actions in the brain and the basis for differential action in men and women: A case for sex-specific medicines. *Pharmacological Reviews, 62*(2), 155–198. http://dx.doi.org/10.1124/pr.109.002071

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology, 102*(4), 700–717. https://doi.org/10.1037/a0026659

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science: a meta-analytic extension of Baye and Monseur (2016). *Large-scale Assessments in Education, 7*(1), 2. https://doi.org/10.1186/s40536-019-0070-9

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Hawken, A., & Munck, G. L. (2013). Cross-national indices with gender-differentiated data: What do they measure? How valid are they? *Social Indicators Research, 111*, 801–838. https://doi.org/10.1007/s11205-012-0035-7

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*(3), 319–334. https://doi.org/10.1177/0022022116687395

Hecht, C. A., Harackiewicz, J. M., Priniski, S. J., Canning, E. A., Tibbetts, Y., & Hyde, J. S. (2019). Promoting persistence in the biological and medical sciences: An expectancy-value approach to intervention. *Journal of Educational Psychology, 111*(8), 1462–1477. https://doi.org/10.1037/edu0000356

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63*(1), 94–105. https://doi.org/10.3102/00346543063001094

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*(2), 150–164. https://doi.org/10.1037/a0015566

Hoffman, R. M., & Borders, L. D. (2001). Twenty-five years after the Bem Sex-Role Inventory: A reassessment and new issues regarding classification variability. *Measurement and Evaluation in Counseling and Development, 34*, 39–55. https://doi.org/10.1080/07481756.2001.12069021

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*(5958), 1410–1412. https://doi.org/10.1126/science.1177067

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592. https://dx.doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist, 74*(2), 171–193. https://doi.org/10.1037/amp0000307

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

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

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences, 106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

Jacobs, J. E., & Bleeker, M. M. (2004). Girls' and boys' developing interests in math and science: Do parents matter? *New Directions for Child and Adolescent Development, 106*, 5–21. https://doi.org/10.1002/cd.113

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., Shefi, N., Pool, J., Urchs, S., Margulies, D. S., Liem, F., Hänggi, J., Jäncke, L., & Assaf, Y. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences, 112*(50), 15468–15473. https://doi.org/10.1073/pnas.1509654112

Joel, D., Tarrasch, R., Berman, Z., Mukamel, M., & Ziv, E. (2014). Queering gender: Studying gender identity in 'normative' individuals. *Psychology & Sexuality, 5*(4), 291–321. https://doi.org/10.1080/19419899.2013.830640

Jones, M. G., Howe, A., & Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. *Science Education, 84*(2), 180–192. https://doi.org/10.1002/(SICI)1098-237X(200003)84:2<180::AID-SCE3>3.0.CO;2-X

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15). University of Nebraska Press.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2003). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 97*(4), 567–583. http://www.jstor.com/stable/3593024

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*(1), 46–66. https://doi.org/10.1093/pan/mpl011

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kyllonen, P. C. (2016). Socio-emotional and self-management variables in learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), The wiley handbook of cognition and assessment (pp. 174–197). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch8

Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–286). CRC Press.

Lenoir, T. (1986). Models and instruments in the development of electrophysiology, 1845-1912. *Historical Studies in the Physical and Biological Sciences, 17*, 1–54. . https://doi.org/10.2307/27757574

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265. https://doi.org/10.1126/science.1261375

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2020). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality,* 1–13. https://doi.org/10.1080/19419899.2020.1729844

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 322*(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education, 32*(5), 669–685. https://doi.org/10.1080/09500690902792385

Mantzicopoulos, P., & Patrick, H. (2010). "The seesaw is a machine that goes up and down": Young children's narrative responses to science-related informational text. *Early Education and Development, 21*(3), 412–444. https://doi.org/10.1080/10409281003701994

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*, 129–149. https://doi.org/10.2307/1163048

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., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 131–198). Academic Press.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59–77. https://doi.org/10.1348/000709910X503501

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meyer, M., Cimpian, A., & Leslie, S. J. (2015). Women are underrepresented in fields where success is believed to require brilliance. *Frontiers in Psychology, 6*, 235. https://doi.org/10.3389/fpsyg.2015.00235

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120*(3), 544–560. https://doi.org/10.1037/a0032459

Morell, V. (1996). Computer culture deflects women and minorities. *Science, 271*(5257), 1915–1916. https://doi.org/10.1126/science.271.5257.1915

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G, Cabrera, H. F., Chen, S. X., Church, A. T., Dougoumalé Cissé, D., … Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*(11), 1423–1436. https://doi.org/10.1177/0146167212451275

OECD (2014). *PISA 2012 technical report.* OECD Publishing.

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41*(3), 181–206. https://doi.org/10.1207/s15326985ep4103_4

Oppermann, E., Keller, L., & Anders, Y. (2020). Geschlechtsunterschiede in der kindlichen MINT-Lernmotivation: Forschungsbefunde zu bestehenden Unterschieden und Einflussfaktoren. *Diskurs Kindheits- und Jugendforschung/Discourse. Journal of Childhood and Adolescence Research, 15*(1), 38–52. https://doi.org/10.3224/diskurs.v15i1.04

Parker, P. D., Van Zanden, B., Marsh, H. W., Owen, K., Duineveld, J. J., & Noetel, M. (2019). The intersection of gender, social class, and cultural context: A Meta-

Analysis. *Educational Psychology Review, 32*, 197–228 https://doi.org/10.1007/s10648-019-09493-1

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology, 114*(S1), 138–170. https://doi.org/10.1016/j.ssresearch.2007.06.012

Popper, K. (2002). *Conjectures and refutations: The growth of scientific knowledge* (2nd ed.). Routledge Academic.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment, 32*(1), 39–51. https://doi.org/10.1027/1015-5759/a000336

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence, 53*, 202–210. https://doi.org/10.1016/j.intell.2015.10.006

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS ONE, 7*(7), e39904. https://doi.org/10.1371/journal.pone.0039904

Reilly, D., Neumann, D. L., & Andrews, G. (2016). Sex and sex-role differences in specific cognitive abilities. *Intelligence, 54*, 147–158. http://dx.doi.org/10.1016/j.intell.2015.12.004

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Investigating gender differences in mathematics and science: Results from the 2011 Trends in Mathematics and Science Survey. *Research in Science Education, 49*(1), 25–50. https://dx.doi.org/10.1007/s11165-017-9630-6 +

Reis, E. (2007). Divergence or disorder? The politics of naming intersex. *Perspectives in Biology and Medicine, 50*(4), 535–543. http://dx.doi.org/10.1353/pbm.2007.0054

Riegle-Crumb, C. (2005). The cross-national context of the gender gap in math and science. In L. Hedges & B. Schneider (Eds.), *The social organization of schooling* (pp. 227–243). Russell Sage Foundation.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*(6), 775–777. https://doi.org/10.1037/0003-066X.45.6.775

Rosenzweig, E. Q., Wigfield, A., & Hulleman, C. S. (2020). More useful or not so bad? Examining the effects of utility value and cost reduction interventions in college physics. *Journal of Educational Psychology, 112*(1), 166–182. https://doi.org/10.1037/edu0000370

Rozek, C. S., Hyde, J. S., Svoboda, R. C., Hulleman, C. S., & Harackiewicz, J. M. (2015). Gender differences in the effects of a utility-value intervention to help parents motivate adolescents in mathematics and science. *Journal of Educational Psychology, 107*(1), 195–206. https://doi.org/10.1037/a0036981

Sahin, A., Gulacar, O., & Stuessy, C. (2015). High school students' perceptions of the effects of international science Olympiad on their STEM career aspirations and twenty-first century skill development. *Research in Science Education, 45*(6), 785–805. https://doi.org/10.1007/s11165-014-9439-5

Schellenberg, D., & Kaiser, A. (2018). The sex/gender distinction: Beyond f and m. In C. B. Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA handbook of the psychology of women: History, theory, and battlegrounds* (p. 165–187). American Psychological Association. https://doi.org/10.1037/0000059-009

Schleicher, A. (2019). *PISA 2018: Insights and interpretations.* OECD Publishing. https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Schmidt, N. &, Fox, K. (2018, December 29). Germany's third gender law is celebrated as a revolution. But some say it's just the first step. *CNN.* https://edition.cnn.com/2018/12/29/health/third-gender-law-germany-grm-intl/index.html

Sczesny, S., Nater, C., & Eagly, A. H. (2019). Agency and communion: Their implications for gender stereotypes and gender identities. In A. E. Abele, & B. Wojciszke (Eds.) *Agency and communion in social psychology. Current issues in social psychology* (pp. 103–116). Routledge.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation, 1993*(60), 13–57. https://doi.org/10.1002/ev.1660

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*, (November), 1–18. https://doi.org/10.2139/ssrn.2694998

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2005). Parents' socializing behavior and children's participation in math, science, and computer out-of-school activities. *Applied Developmental Science, 9*(1), 14–30. https://doi.org/10.1207/s1532480xads0901_3

Smith, J. L., Lewis, K. L., Hawthorne, L., & Hodges, S. D. (2013). When trying hard isn't natural: Women's belonging with and motivation for male-dominated STEM fields as a function of effort expenditure concerns. *Personality and Social Psychology Bulletin, 39*(2), 131–143. https://doi.org/10.1177/0146167212468332

Spence, J. T., & Helmreich, R. (1978). *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents.* University of Texas.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sticca, F., Goetz, T., Nett, U. E., Hubbard, K., & Haag, L. (2017). Short-and long-term effects of over-reporting of grades on academic self-concept and achievement. *Journal of Educational Psychology, 109*(6), 842–854. https://doi.org/10.1037/edu0000174

Stryker, S. (2008). *Transgender history.* Seal Press.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS ONE, 8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence, 48*, 137–151. https://doi.org/10.1016/j.intell.2014.11.006

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Stoet, G., & Geary, D. C. (2020a). Sex-specific academic ability and attitude patterns in students across developed countries. *Intelligence, 81*, 101453. https://doi.org/10.1016/j.intell.2020.101453

Stoet, G., & Geary, D. C. (2020b). The gender-equality paradox is part of a bigger phenomenon: Reply to Richardson and colleagues (2020). *Psychological Science, 31*(3), 342-344. https://doi.org/10.1177/0956797620904134

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PloS ONE, 11*(3), e0150194. https://doi.org/10.1371/journal. pone.0150194

Storage, D., Charlesworth, T. E. S., Banaji, M. R., & Cimpian, A. (2020). Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology, 90,* 104020. https://doi.org/10.1016/j.jesp.2020.104020

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6*, 189. https://doi.org/10.3389/fpsyg.2015.00189

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

Tate, C. C., Youssef, C. P., & Bettergarcia, J. N. (2014). Integrating the study of transgender spectrum and cisgender experiences of self-categorization from a personality perspective. *Review of General Psychology, 18*, 302–312. http://dx.doi.org/10.1037/gpr0000019

Tenenbaum, H. R., & Leaper, C. (2003). Parent-child conversations about science: The socialization of gender inequities? *Developmental Psychology, 39*(1), 34–47. https://doi.org/10.1037/0012-1649.39.1.34

Vonkova, H., Bendl, S., & Papajoanu, O. (2017). How students report dishonest behavior in school: Self-assessment and anchoring vignettes. *The Journal of Experimental Education, 85*(1), 36–53. https://doi.org/10.1080/00220973.2015.1094438

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In A. J. Elliot, *Advances in Motivation Science* (Vol. 7, pp. 161-198). https://doi.org/10.1016/bs.adms.2019.05.002

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In J. M. Olson & M. P. Zanna, *Advances in experimental social psychology* (Vol. 46, pp. 55–123). Academic Press.

Wood, W., & Eagly, A. H. (2015). Two traditions of research on gender identity. *Sex Roles, 73*, 461–473. http://dx.doi.org/10.1007/s11199-015-0480-2

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist, 70*(1), 10–20. https://doi.org/10.1037/a0038208

Erklärung

Hiermit versichere ich, dass ich die Dissertation „The Interplay of Achievement and Achievement Motivation: Gender Differences in Math Top-Performers and Functional Relations“ selbständig verfasst habe. Alle Hilfsmittel, die ich verwendet habe, sind angegeben. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder angelehnt worden.

Berlin, Juni 2020

_____

Lena Kristina Keller

Eigenanteil und Veröffentlichungen

Die folgende Tabelle veranschaulicht den Eigenanteil an den veröffentlichten oder zur Veröffentlichung eingereichten wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift.

Autoren	Titel	Status	Eigenanteil
Keller, L., Preckel, F., Eccles, J. S., & Brunner, M.	Top-performing math students in 82 countries: A meta-analysis of gender differences in achievement, achievement profiles, and achievement motivation	In Begutachtung in *Review of Educational Research*	Konzeption der Fragestellung (überwiegend), Aufarbeitung der Literatur und des theoretischen Hintergrunds (vollständig); Datenaufbereitung (vollständig), Datenanalyse (vollständig); Verfassung des Manuskriptes (überwiegend), Antworten auf Gutachten (überwiegend)
Keller, L., Preckel, F., & Brunner, M.	Nonlinear relations between achievement and academic self-concepts in elementary and secondary school: An integrative data analysis across 13 countries.	Zur Veröffentlichung angenommen in *Journal of Educational Psychology*	Konzeption der Fragestellung (überwiegend), Aufarbeitung der Literatur und des theoretischen Hintergrunds (vollständig); Datenaufbereitung (vollständig), Datenanalyse (vollständig); Verfassung des Manuskriptes (überwiegend), Antworten auf Gutachten (überwiegend)

Berlin, Juni 2020

Lena Kristina Keller

For reasons of data protection, the curriculum vitae is not published in the electronic version.