The Dimensional Structure of Students' Self-Concept and Interest in Science Depends on Course Composition

Malte Jansen,

Institute for Educational Quality Improvement, Berlin, Germany

Centre for International Student Assessment, Germany

Ulrich Schroeders

Department of Psychology, University of Kassel

Oliver Lüdtke

Leibniz Institute for Science and Mathematics Education, Kiel

Centre for International Student Assessment, Germany

Herbert W. Marsh

Australian Catholic University, Sydney, Australia

Author Note

Accepted manuscript.DOI: 10.1016/j.learninstruc.2018.11.001.

Malte Jansen, Institute for Educational Quality Improvement (IQB), Berlin, Germany, and Centre for International Student Assessment (ZIB), Germany; Ulrich Schroeders, Department of Psychology, University of Kassel, Germany; Oliver Lüdtke, Leibniz Institute for Science and Mathematics Education, Kiel, Germany, and Centre for International Student Assessment (ZIB), Germany; Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University, Stratfield Campus, Australia.

Correspondence concerning this article should be addressed to Malte Jansen, Humboldt University Berlin, Institute for Educational Quality Improvement, Unter den Linden 6, 10099 Berlin, Germany, Email: Malte.Jansen@IQB.HU-Berlin.de, Telephone: +04930209346587

Further Notes

We thank the research data center (FDZ) at the Institute for Educational Quality Improvement (IQB) for providing data access. This research was partly funded by the Center for International Student Assessment (ZIB).

This manuscript is a widely extended version of a short report published in German:

Jansen, M., Schroeders, U., Lüdtke, O., & Pant, H. A. (2014). Interdisziplinäre Beschulung und die Struktur des akademischen Selbstkonzepts in den naturwissenschaftlichen Fächern. [Interdisciplinary teaching and the structure of academic self-concept in the sciences] *Zeitschrift für Pädagogische Psychologie*, 28(1), 43–49. https://doi.org/10.1024/1010-0652/a000120

Among other changes, we extended the analyses to include more than two science teaching groups, examine gender differences (using both MGCFA and MIMIC models) included interest in addition to self-concept and introduced several robustness checks including propensity score matching.

Abstract

Both academic self-concept and interest are considered domain-specific constructs. Previous research has not yet explored how the composition of the courses affects the domain-specificity of these constructs. Using data from a large-scale study in Germany, we compared ninth-grade students who were taught science as an integrated subject with students who were taught biology, chemistry, and physics separately with regard to the dimensional structure of their self-concepts and interests. Whereas the structure of the constructs was six-dimensional in both groups (self-concept and interest factors for biology, chemistry, and physics), the correlations between the domain-specific factors were higher in the integrated group. Furthermore, the pattern of gender differences differed across groups. Whereas male students generally showed higher self-concept and interest in physics and chemistry, a small advantage for male students in biology was only present in integrated science teaching group. We conclude that aspects of the learning environment such as course composition may affect the dimensional structure of motivational constructs.

Keywords: academic self-concept, interest, integrated science teaching, dimensionality, construct differentiation

The Dimensional Structure of Students' Self-Concept and Interest in Science Depends on Course Composition

1 Introduction

Academic self-concepts and academic interests are understood as domain-specific constructs; that is, students feel more confident and interested in some academic domains than in others. This theoretically assumed domain-specificity has been practically operationalized by measuring self-concept and interest separately for different school subjects (e.g., mathematics self-concept, English self-concept, etc.). However, school subjects vary by grade level, school type, and country. For example, science domains such as physics and biology can be taught either as part of an integrated science class or as separate school subjects. Researchers have yet to explore how such organizational characteristics affect the dimensional structure of domain-specific self-concepts and interests.

In this study, we make use of a natural experiment in Germany to close this research gap. Following the approach in other countries, integrated science teaching instead of the traditional approach of teaching biology, chemistry, and physics separately is now increasingly offered in Germany. Using a nationally representative dataset of ninth-grade students, we analyzed self-concept and interest measures in biology, chemistry, and physics. We compared students who were taught these domains as separate school subjects with students who were taught science as an integrated subject regarding (a) the dimensional structure of their academic self-concept and interest as well as (b) the pattern of gender differences in self-concepts and interests. We expected the students who were taught science as an integrated subject to show a less differentiated factor structure for both self-concept and interest.

2 Theoretical Background and State of Research

2.1 The Structure of Academic Self-Concept: Development and Differentiation

Shavelson, Hubner, and Stanton (1976) described academic self-concept as a multidimensional and domain-specific construct (e.g., academic self-concept in English, in mathematics). Since then, a plethora of empirical research has shown that such domain-specific facets can be distinguished (Brunner et al., 2010; Marsh, 1990). In the context of school education, the abovementioned domain-specificity has been operationalized by measuring self-concepts in different school subjects. However, which science subjects are taught in a given learning environment varies substantially: Whereas in some countries such as the US the focus is on integrated science teaching, in other countries such as Germany physics, chemistry, and biology are usually taught separately. Accordingly, some educational large-scale studies (e.g., the PISA studies) have examined students' self-concepts in general science, whereas others have used more differentiated measures related to the science subdomains. Such research has shown that students differentiate between their academic self-concepts in biology, chemistry, and physics and that these facets are only moderately related (Hardy, 2014; Jansen, Schroeders, & Lüdtke, 2014).

Comparison processes based on achievement feedback are considered an important source of students' academic self-concept (Marsh et al., 2018; Wolff, Helm, Zimmermann, Nagy, & Möller, 2018). Students compare their current achievement in a domain with the achievement of their peers in the same domain (social comparisons), their own past achievement in the same domain (temporal comparisons), and their own achievement in other domains (dimensional comparisons). These comparison processes can explain individual differences in academic self-concept as well as the finding that students with similar achievement have diverging domain-specific self-concepts (Retelsdorf, Köller, & Möller, 2014; Wolff et al., 2018). However, comparison processes have mainly been studied

with regard to their effect on the mean level of self-concepts rather than the dimensional structure of self-concept.

It has been suggested that the structure of self-concept should become more differentiated over the course of students' school careers offering as students receive more domain-specific achievement feedback (Shavelson et al., 1976; Stipek & Mac Iver, 1989). This idea is consistent with empirical evidence that different self-concepts tend to become more reliable and distinct with age (Marsh & Ayotte, 2003). Nevertheless, already primary school students in Grade 1 have been found to differentiate between a mathematical and verbal self-concept that become even more differentiated across primary school (Ehm, Lindberg, & Hasselhorn, 2014; Marsh & Ayotte, 2003; Schmidt et al., 2017). In contrast, Arens and Morin (2016) did not find differences in the correlational pattern between math and German across Grades 3 to 6, thus, hypothesizing that, for these two specific domains, the "differentiation process is already completed" (p. 22). In general, there is little evidence for further differentiation of academic self-concept after primary school, especially in the domain of science. The question which role structural characteristics such as course composition plays in the development of the structure of domain-specific self-concept has yet to be answered.

2.2 The Structure of Academic Interests: Development and Differentiation

According to expectancy-value theory (Wigfield & Eccles, 2000), achievement motivation is affected not only by self-beliefs about one's competence (i.e., what students think they can do) but also by value beliefs (i.e., what students like to do). The most prototypical value belief is intrinsic value or individual *interest*. A student with high individual interest would be characterized by a strong cognitive commitment and emotional attachment to a specific domain (Krapp, 2002). Thus, individual interest is assumed to be inherently domain-specific (Hidi & Renninger, 2006; Krapp, 2002).

Compared with research on academic self-concepts, structural models of academic interest across academic domains have been less prominent. Results from previous research in an educational context confirmed the assumption of domain-specificity and, in the domain of science, have also suggested multidimensionality (Jansen, Lüdtke, & Schroeders, 2016): Similar to self-concepts, students held different levels of interest in biology, chemistry, and physics with the latter two more closely related. That the factor structure seems to be similar to the structure of academic self-concepts is not surprising given that the two constructs are substantially correlated and have been shown to co-evolve during the school career (Archambault, Eccles, & Vida, 2010; Denissen, Zarrett, & Eccles, 2007). Thus, they also show quite similar empirical patterns in their relations to other constructs even though the relation between achievement and self-concept is stronger than relation between achievement and interest (Trautwein et al., 2012).

Even though the constructs are empirically strongly intertwined and are assumed to co-predict achievement motivation following expectancy-value-theory (Wigfield & Eccles, 2000), their theoretical foundations differ: While the idea of achievement feedback and comparison processes as sources of interest has been less prominent (for an exception, see Schurtz, Pfost, Nagengast, & Artelt, 2014), more emphasis was put on the developmental part of individual interest from initial situational interest (e.g., the Four-Phase Model of interest development by Hidi and Renninger 2006). One key assumption of this model is that the structure of interest becomes more elaborated over time and that interest becomes more differentiated across a student's school career. Hidi and Renninger also emphasized the importance of the learning environment for providing opportunities for students to engage with the content because individual interest can develop only through continuous engagement. But still, the effect of course composition which sets the learning environment on interest development has not been studied so far.

2.3 The "Gender Gap" in Science Motivation

Gender differences in educational choices related to the STEM domains (Science, Technology, Engineering, and Mathematics) are still a matter of concern in educational systems worldwide (OECD, 2015). Compared with men, women are less likely to pursue higher education in STEM fields (Wang, Eccles, & Kenny, 2013) even though no substantial gender differences have been found in mathematics and science achievement (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). A closer look, however, reveals differences across scientific domains (Cheryan, Ziegler, Montoya, & Jiang, 2017). Even given comparable school achievement, only 22% of first-year university physics students in Germany are female, compared with 43% female first-year students in chemistry and even 62% in biology (Schroeders, Penk, Jansen, & Pant, 2013). In the same vein, Wang and Degol (2013, p.20) concluded in a recent review examining that "little of this [previous] work has focused on the different occupational choices within STEM (e.g., physical sciences versus biological sciences)".

Expectancy-value theory proposes differences in academic self-concepts and value beliefs as a central driver of differences in educational choices (Eccles, 2011; Wang & Degol, 2013). This assumption has been strengthened by studies showing that self-concepts and value beliefs affect both aspirations in relation to a career in a STEM field (Guo, Marsh, Parker, Morin, & Dicke, 2017; Nagengast et al., 2011) and actual educational choices (Parker et al., 2012). There is also evidence for gender differences in these facets of motivation. Advantages for male students in self-concepts and interests in mathematics have consistently been reported to emerge as early as the beginning of elementary school (Eccles, Wigfield, Harold, & Blumenfeld, 1993; Hyde, Fennema, Ryan, Frost, & Hopp, 1990; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). In science, the pattern is more complex: There is

a strong advantage for male students in physics, a smaller advantage in chemistry, and no gender difference in biology (Elster, 2007; Hardy, 2014; Jansen et al., 2014).

2.4 Integrated versus Separated Science Teaching and its Effect on the Factorial Structure of Self-Concept and Interest

In Germany, science education in secondary school has a long tradition of separately teaching science subjects. Usually, biology is the first science subject taught from Grade 5 onwards, followed by physics and chemistry in Grades 7 and 8, respectively. To promote general scientific literacy, integrated science teaching is now increasingly offered in Germany with the aim of emphasizing relations and commonalities between the different domains. In most cases, these integrated science subjects are offered in nonacademic track schools (that is, all school types other than *Gymnasium*; see Supplement A).[1] But how does the teaching approach affect the development of the structure of self-concept and interest? Several potential mechanisms are conceivable:

First, the classes the students attend are labeled differently ("science" vs. "biology," "chemistry," "physics") even though the general curriculum is quite similar (i.e., based on the national educational standards). For example, having 2 lessons of "biology," 2 lessons of "chemistry," and 2 lessons of "physics" each week rather than 6 lessons of "science" could have led students to develop separate motivational dispositions if these were indeed strongly aligned with the offered school subjects. The labels could, for example, trigger different stereotypes of students who do well in these subjects. Previous studies show that students develop such school-subject specific stereotypes and that their domain-specific self-concepts are influenced by self-prototype matching processes (Hannover & Kessels, 2004; Kessels, 2005)

[1] Sometimes, integrated science is also offered in addition to domain-specific teaching or only in earlier grades (for a more detailed overview, see Czerniak & Johnson, 2011; Executive Agency Education, Audiovisual and Culture, 2011).

Second, the school subject structure is also mirrored in grading. Students who are taught biology, chemistry, and physics separately also receive separate grades in each subject, whereas, in most cases, students who are taught integrated science receive only one grade. As grades are a very salient and important source of achievement feedback for students and this feedback, interpreted through comparison processes, is the base for students' (Jansen, Schroeders, Lüdtke, & Marsh, 2015; Möller & Marsh, 2013; Wolff et al., 2018), we would expect the structure of self-concept facets to mirror the structure of school subjects. More specifically, students compare and contrast their achievement across different domains if there is more than one facet of achievement feedback (dimensional comparisons; Möller & Marsh, 2013) as shown not only in a plethora of observational studies (Möller, Pohlmann, Köller, & Marsh, 2009), but also experimental studies based on manipulated achievement feedback (Möller & Köller, 2001; Strickhouser & Zell, 2015).. Within the domain of science, such contrast effects have been shown between biology and physics (Jansen et al., 2015). The strength of contrast or assimilation effects variy depending on the perceived similarity between two subjects. Within the context of course composition, we think that the tailored achievement feedback that occurs when science domains are taught as separate subjects, will lead students to increasingly contrast their performance in those domains. In turn, this will result in a more differentiated factor structure with lower correlations between self-concept facets although their performance might be highly correlated (Jansen et al., 2014, 2015). Since dimensional comparison effects have also been found for interest (Schurtz et al., 2014) and students who feel competent also shown higher interest (Archambault et al., 2010), we would also expect a more differentiated interest structure to result from more differentiated grading.

Third, an integrated science subject is taught by a single teacher, whereas students are usually taught by different teachers who pursued different subject-specific courses in their

university education. Teachers will differ in their style and quality of instruction, their grading, and their student support, et cetera. All this is likely to influence students' self-concepts and interests (Kunter et al., 2013; Opdenakker, Maulana, & Brok, 2012). Subsequently, if the domains at hand are taught by different teachers, this might contribute to a more differentiated, domain-specific structure of these constructs. More specifically, different teachers might provide different opportunities for interest to develop, for example, by providing different levels of cognitive activation and support (Lazarides & Ittel, 2012; Tsai, Kunter, Lüdtke, Trautwein, & Ryan, 2008), which might result in a more differentiated structure of interest.

Fourth, in addition to these more general mechanisms that are independent of the investigated domains, specific characteristics of integrated science teaching could lead to a less differentiated structure of self-concept and interest. For example, in integrated classes, there is a stronger focus on domain-general competencies such as scientific literacy, problem-solving skills, and the methods of scientific inquiry (Chowdhary, Liu, Yerrick, Smith, & Grant, 2014; Czerniak & Johnson, 2011). Moreover, integrated science teaching emphasizes the commonalities between the three science domains, often exemplified with real-world problems. For example, one popular method of integrated science teaching is to focus on one topic with a clear reference to everyday life and then discuss aspects of this topic from the perspectives of biology, chemistry, and physics. It should be noted again, however, that the national educational standards for science in secondary school are valid for all federal states independent of the composition of science courses and that generic competencies and including real-world problems is a pivotal goal of science teaching irrespective of the approach.

All four mechanisms point into the direction of stronger construct differentiation when science is taught in separate subjects. However, does this imply that there would be no

differentiation at all for students attending integrated science classes? In general, people can possess self-concepts about areas for which there is no scholarly achievement feedback and that have no relation to academic domains (e.g., social self-concept or physical self-concept; Shavelson et al., 1976). In the academic domain, students also have distinguishable self-concepts about their abilities in the subdomains of school subjects. For example, they differentiate between reading, writing, listening, and speaking self-concepts in language domains, even though these skills are strongly related on an achievement level (Arens & Jansen, 2016). In summary, school subjects are not the least divisible unit of self-concepts but can be studied on a more fine-grained level. Contrary to self-concept research, interest research has more often moved beyond the level of school subjects to the level of more specific topics within domains such as science, and researchers have suggested that a distinction be made between interest in a school subject and interest in a particular topic—for example, students could be generally interested in topics in physics classes such as optics or astronomy but still show low interest in physics as a school subject as taught at a given time (Häussler & Hoffmann, 2000; Krapp & Prenzel, 2011). Thus, we would expect that students can differentiate between their self-concepts and interests in topics in the field of biology, chemistry and physics, even if they are taught within a single subject.

3 The Present Study

Using a nationally representative data set of German ninth-graders, we compared students who were taught science as an integrated subject from Grades 5 to 9 (integrated science teaching; abbreviated as “IST” in the following) with students who were taught biology, chemistry, and physics as separate school subjects (separated science teaching; abbreviated as “SST” in the following) with regard to the structure of their self-concepts and interests. Following the mechanisms outlined in the introduction, we hypothesized that a six-dimensional structure (self-concept and interest factors in biology, chemistry, and physics)

would be found in both teaching approaches. However, we expected the self-concept and interest structures of SST students to be more differentiated compared with the IST students. If the proposed mechanisms were indeed at work and resulted in stronger differentiation in the SST group, the factor correlations between the self-concept and interest facets in biology, chemistry, and physics should be higher in the IST group compared with the SST group.

As a second strand of analysis, we studied gender differences. We expected to find a pattern of stereotypical gender differences in both groups as found in previous studies (Hardy, 2014; Jansen et al., 2014), that is, male advantages should be strongest in physics and still substantial in chemistry, whereas there should be no differences or small advantages for female students in biology.

Given our hypothesis that the self-concept and interest factors should be more differentiated when the domains are taught as separate school subjects, we would additionally expect that more differentiated factors would also go along with a stronger differentiation of gender differences. That is because, lower factor correlations would also allow for stronger differences in the relation between domain specific self-concept and interest factors and other variables such as gender. Based on this methodological argument, we would not expect a mean shift in the gender differences (i.e., the overall differences across the three domains becoming smaller or larger). In addition, it could be argued that characteristics of IST could lead to a mean shift and thus might help reduce gender differences. For example, teachers could more easily place a physical problem in a biological context and thereby raise female students' interest in physics by tapping into their well-established topic-level interest in human biology and the environment (Beier & Ackerman, 2003; Elster, 2007). However, theoretically, such "contagion" effects could also work in the opposite direction and there are no previous studies on the effects of IST on gender differences that could inform a directed hypothesis.

4 Method

4.1 Study Design

The data were derived from a large-scale assessment study focusing on students' proficiency in math and science at the end of secondary education (IQB National Assessment Study 2012, see Lenski et al., 2016; Pant et al., 2015). The data set is one of the largest school assessment data sets available in Germany and consists among others of a students' and school principals' questionnaire. To determine the way science was taught, the principals of all participating schools were asked to specify how many weekly lessons of biology, chemistry, physics, and integrated science, respectively, were taught in each semester from Grade 5 to Grade 9. The principal questionnaire was mandatory for some of the federal states but not for all of them. A total of 890 school principals (67%) participated.

Participation was mandatory for the randomly sampled public schools, thus enabling a participation rate of 96.7% at the school level and 92% at the student level. The total sample consisted of 44,584 ninth-grade students from 1,326 schools. The students were on average 15.02 years old (SD = 0.66); about half of them were female (49.5%). As is common in large-scale assessment, different test booklets and student questionnaires were randomly assigned to students (planned missing data design; Little, Jorgensen, Lang, & Moore, 2014; Rhemtulla & Little, 2012). The data were collected in spring 2012.

4.2 Measures

Academic self-concept. We assessed domain-specific self-concepts in biology, chemistry, and physics with four similarly worded items per domain: "I learn quickly in [biology/chemistry/physics]," "I have always believed that [biology/chemistry/physics] is one of my best subjects," "I get good grades in [biology/chemistry/physics]," and "I am just not good at [biology/chemistry/physics]" (reverse scored). Students replied on a 4-point scale that had the response options *strongly agree* (coded as 4), *agree* (3), *disagree* (2), and

strongly disagree (1). The items were taken from the international questionnaire used in PISA 2003 (originally for measuring self-concept in mathematics). The version adapted for the three science domains was also used in a previous study (Jansen et al., 2015). The second and the third items may have been more difficult for the IST group to answer (because students usually received only one grade, and there were no separate subjects). However, the self-concept scales showed good reliability in both groups in our sample (coefficient $\omega > .85$, see Table 1), and they demonstrated measurement invariance across the two groups (see Table 1).

Interest. Interest was measured in biology, chemistry, and physics with items that had parallel wording: "I am interested in [biology/chemistry/physics]," "[biology/chemistry/physics] is important for me personally," "Studying [biology/chemistry/physics] is fun," and "Studying [biology/chemistry/physics] is one of my favorite activities. The response options, which used a 4-point scale, were *strongly agree* (coded as 4), *agree* (3), *disagree* (2), and *strongly disagree* (1). The items were taken from the German national questionnaire from the PISA studies (Ramm, Adamsen, Neubrand, & Deutsches PISA-Konsortium, 2006) and originally addressed interest in mathematics. The items refer to the cognitive and affective components of habitual, domain-specific interest (Krapp & Prenzel, 2011). The items were also used in a previous study on interest (Jansen et al., 2016). The reliability coefficients were very high in both groups ($\omega > .91$, see Table 1).

4.3 Sample of Analysis and Treatment of Missing Data

The total sample of the *IQB National Assessment Study 2012* consisted of 44,584 students. Using the data from the principal questionnaire and the participation rate of 67% mentioned above, there was information about the number of science lessons for only 32,512 students; all other students were excluded. Of these students, 13,307 received a version of the questionnaire that included self-concept items; 8,040 questionnaires also included items

about domain-specific interest. This was because the self-concept items were included in three of the eight versions of the questionnaires that were used, whereas interest was included in only two. Such planned missing data designs are often used in educational large-scale assessment as an economic and efficient way to collect data in a reasonable amount of time. The resulting dataset has data *Missing Completely at Random* (MCAR) which can easily be addressed with statistical techniques and does not induce bias (e.g., Enders, 2010; Little et al., 2014). For students who received a questionnaire that included the self-concept and interest items, respectively, the item-level rate of missing data was very low, ranging from 1.0% to 2.3%.

Because we aimed to analyze self-concept and interest jointly in a single model, we included all students who completed at least one self-concept item in the sample (n = 12,967). From this sample, students were further assigned to the SST or IST group. As an additional robustness check, we also replicated the results with other sample selection procedures (see Supplement G).

We estimated all models using *Full Information Maximum Likelihood* (FIML) estimation—a model-based approach for handling missing data. In the FIML procedure, missing data and parameter estimation are combined in a single step (Enders, 2010). FIML is considered superior to traditional methods of treating missing data because it allows for more precise parameter estimation with higher statistical power (Schafer & Graham, 2002). Furthermore, because it is unbiased under the MCAR and *Missing At Random* (MAR) assumptions, it is appropriate for planned missing data resulting from randomly distributed versions of a questionnaire.

Integrated teaching sample (IST). Students were assigned to the integrated teaching sample if they had classes in integrated science from Grade 5 through Grade 9 and no classes in any of the separate science subjects during this period. As mentioned above, the

form of integrated science teaching varies, and a mixed approach of integrated and separate science teaching is also common. However, not only would the many different forms have been hard to disentangle in our data set, but any mixed forms would also have made it harder to interpret possible group differences. Therefore, we decided to use the abovementioned strict criterion for the integrated teaching sample (a robustness check that included students with mixed forms of science teaching showed their result patterns were largely similar to the SST group, see Supplement E). This criterion was fulfilled by 337 students (of which 177 students completed both the self-concept and interest items, and 160 students completed only the self-concept items). As expected given the practice of integrated science teaching in Germany, all IST students were from nonacademic track schools.

Separate science teaching sample (SST). Because only students from the nonacademic track could be identified for the IST sample, we also excluded all academic track students to ensure comparability. We assigned students to the separate teaching group if they had never received instruction in an integrated science subject from Grade 5 to Grade 9 and were thus taught biology, chemistry, and physics only as separate school subjects. This did not imply that SST students took every subject every semester but that they took each of the subjects at some time during secondary school. On the basis of this definition, we identified 4,361 SST students.

4.4 Data Analysis

Model estimation. We used a multigroup CFA framework to examine our central research question—whether the factor structure in the IST group would be less differentiated compared with the SST group. In both groups, the measurement model included six correlated first-order factors representing the academic self-concepts and interests in biology, chemistry, and physics. Furthermore, we tested whether there were differential mean gender differences in the IST versus the SST group. To be able to test these parameter differences

(correlations, means), a series of measurement invariance tests were conducted. The procedure is described in detail in Supplement B. All models were estimated using the software Mplus 7.1. Correlated residuals between parallel worded items across domains were estimated in all models (e.g., “I learn quickly in biology” and “I learn quickly in chemistry”). To account for the hierarchical structure of the data (students nested within classes), we corrected the standard errors (TYPE = COMPLEX). In order to check and ensure the comparability between the IST and the SST group, we also replicated all analyses using matched samples (see Supplement F).

5 Results

5.1 Invariance Tests and Factor Correlations

After first examining descriptive statistics (see Supplement C), we tested our measurement model (six first-order factors representing self-concept and interest in the three domains) for measurement invariance across the SST and IST groups. The configural invariance model showed a good fit to the data (CFI = .983, RMSEA = .030, SRMR = .028; see Table 1, Model 1). On the basis of this model, we consecutively fixed the factor loadings (weak invariance; Table 1, Model 2), the factor loadings and intercepts (strong invariance, Table 1, Model 3), and the factor loadings, intercepts, and residual variances (strict invariance, Table 1, Model 4) to be equal across groups. Even for the strict level of invariance, the decrease in model fit was very small (configural vs. strict invariance: Δ CFI = -.001, Δ RMSEA = -.001), indicating that strict measurement invariance could be established.

In the next step, we used the strict invariance model to further test for invariance in the factor variances and covariances using the χ^2-difference test. The factor variances could be constrained to equality without a significant decrease in model fit ($\Delta\chi^2 = 6.92$, $df = 6$, $p = .33$; see Table 1, Model 5). However, when the factor covariances were also constrained to

equality, the χ^2-difference test turned out to be significant ($\Delta\chi^2 = 136.41$, $df = 15$, $p < .01$, see Table 1, Model 6).

Thus, as expected, there were significant and substantial differences in the pattern of correlations (see Table 2). All six constructs showed positive correlations in both groups. The correlations were higher within each domain (e.g., biology self-concept and interest in biology showed higher relations than biology self-concept and interest in chemistry) and within each construct (e.g., interest in biology and interest in chemistry showed higher relations than interest in biology and chemistry self-concept). For both self-concept and interest, the correlations between the domain-specific measures in biology, chemistry, and physics were much stronger in the IST group compared with the SST group as expected. That is, the correlations between the three self-concept facets ranged from .71 to .82 in the IST group, whereas the facets were only moderately correlated in the SST group (r = .30 to 45). This pattern was similar for interest with high relations between interest in biology, chemistry, and physics in the IST group (r = .66 to .87) and substantially lower relations in the SST group (r = .30 to .54).

5.2 Gender Differences

After first establishing strong measurement invariance (for a description of the procedure, see Supplement B; for the model fit results, see Supplement D), we compared the factor means and the effect size (Cohen's d) of the gender differences for the IST and SST groups. The results are shown in Figure 1. In the SST group, we found a pattern of stereotypical gender differences that replicated previous results (e.g., Jansen et al., 2014): There were strong advantages for male students in physics (self-concept: d = 0.60, interest: d = 0.71), smaller but still substantial advantages for male students in chemistry (self-concept: d = 0.31, interest: d = 0.32), and only small differences in favor of female students in biology (self-concept: d = -0.10, interest: d = -0.18). In the IST group, there were still

strong advantages for male students in physics (self-concept: $d = 0.56$, interest: $d = 0.52$), advantages in chemistry (self-concept: $d = 0.45$, interest: $d = 0.44$) and, most interestingly, also substantial advantages for male students in biology (self-concept: $d = 0.28$, interest: $d = 0.22$). A z-test revealed that only the effect size of the gender difference in biology (self-concept: $z = 2.853$, $p < .01$; interest: $z = 2.946$, $p < .01$) but not in chemistry and physics differed between the IST and SST groups. As an additional test of the robustness of this difference, we estimated a multiple indicator multiple cause (MIMIC) model instead of using a multigroup approach finding similar results (see Supplement H).

6 Discussion

We examined the effect of the course composition on the structure of students' academic self-concept and interest in the sciences. Whereas both groups showed a domain-specific structure, the correlations between the three domain-specific self-concept and interest factors (biology, chemistry, physics) were considerably higher in the group that had received integrated science teaching instead of being taught biology, chemistry, and physics as separate school subjects. Furthermore, the pattern of gender differences in self-concept and interest differed between the groups. Whereas male students showed higher self-concept and interest than female students in physics and chemistry in both groups, their interest and self-concept in biology were marginally lower compared to female students in the SST group, but higher in the IST group. The results were very similar for academic self-concept and interest. Overall, the results suggest that course composition (i.e., the structure of school subjects that are taught) might significantly influence the differentiation of academic self-concepts and interests.

6.1 Implications for Research on Self-Concept and Interest

We believe these results contribute to the literature on the structure of self-concepts and interest in several ways. Using the natural experiment of IST versus SST as a case study

is an innovative way to test effects of structural characteristics of the learning environment more broadly, and the composition of a course more specifically on the development of not only the means but the factor structures of motivational constructs. It has long been argued that domain specificity is a key characteristic of both self-concept and interest (Brunner et al., 2010; Hidi & Renninger, 2006; Krapp & Prenzel, 2011; Marsh, 1990; Shavelson et al., 1976). Our results show the central role of school subjects in shaping that domain specificity. Academic self-concepts and interest seem to be linked to the specific ways in which curricula (i.e., differentiation of courses) and achievement feedback (i.e., differentiation of grading) are structured within schools. We outlined four—potentially interacting and cumulating—explanations for this effect: different course labels, differential grading, different teachers, and specific characteristics of integrated science teaching. Unfortunately, these cannot be tested in our study, but as will be outlined below, should be the focus of further studies. Still, we argue that first describing the phenomenon of course composition effects on the structure of motivational construct is an important first notion, especially since self-concept and interest are highly popular constructs in educational psychology and their domain-specificity has received so much attention in previous research.

Overall, we think the results point to the issue that even though the terms (school) "subject" and "domain" (e.g., "subject-specific self-concept" or "domain-specific self-concept") have largely been used interchangeably in the self-concept literature, they might mean different things in different school systems. Whereas the same domains (biology, chemistry, physics) were covered in both groups, the allocation of these domains to school subjects (one vs. three subjects) differed, and this difference seems to have had a strong effect even though the curriculum was quite similar between the groups. With regard to interest rather than self-concept, this idea has already been formulated. For example, in their comprehensive review of interest in science, Krapp and Prenzel (2011) described, "In order

to distinguish between different kinds of science interests, it is obvious to refer to the structure of school subjects because these mainly provide the opportunities to get in touch with sciences systematically" (p. 33). However, they also pointed out that students can still be interested in different topics, contexts, and activities within and across school subjects. Based on this idea, more specific models have been developed to represent domain-specific interest (e.g., including many facets of interest in physics measured by as many as 88 items; Häussler & Hoffmann, 2000). Thus, in the interest literature, the level of school subjects has been described as only one of many levels that can be examined, whereas this point has not been stressed as much in the self-concept literature.

On a more general stance, our results show the importance of the learning environment for students' responses to questionnaire items assessing self-concept and interest . One recommendation for future studies, and especially for cross-cultural studies, is that researchers should pay attention to the characteristics of the learning environment when studying domain-specific self-concepts and interests. For example, can students choose their courses and course levels? Or do all students receive the same amount of teaching? Is there implicit or explicit tracking (Chmielewski, Dumont, & Trautwein, 2013)? Can students choose not to attend courses in a given subject, and if so, how would the answers students give to self-concept or interest items be interpreted in that domain? How exactly are the courses labeled, and how do these labels compare with the items that are used? The answer to these more organizational questions which have been mostly neglected in previous research could have an influence on the structure of school-related psychological constructs. Science education is probably the most prominent example to what extent course composition differ across and even within countries. However, there are other situations in which course composition might play a crucial role. For example, in the US, math teaching has traditionally been strongly separated by topics such that entire courses of algebra,

geometry, and pre-calculus are taught separately in separate school years. In most other countries, there is one integrated mathematics subject, and the topics change at a faster pace within each school year. The first case would be more likely to result in the students developing separate self-concepts or interests in different subdomains of math such as algebra and geometry compared with the second case.

6.2 Gender Differences

Female students typically show lower self-concepts and interests in science than male students do, particularly in chemistry and physics (Jansen et al., 2014). We replicated these results in our study. However, our focus was on comparing the pattern of gender differences in the IST and SST groups. We expected a less differentiated pattern of gender differences in the IST group compared with the SST group. On a descriptive level, whereas the effect sizes for the gender differences in the three domains were more similar in the IST group than in the SST group, this pattern was mostly driven by self-concept and interest in biology. In this domain, there were no gender differences in the SST group (a finding that is consistent with previous research on academic self-concepts; Hardy, 2014; Jansen et al., 2014). However, in the IST group, male students showed an unexpected advantage in biology. The gender differences in chemistry and physics were not statistically different across the course composition groups.

Regarding the discussion about gender gaps in STEM career choices, it seems difficult to decide whether a less differentiated factor structure of self-concept and interest and, correspondingly, a less differentiated structure of gender differences, would be desirable from a normative perspective. In our study, there was no significant reduction in female students' disadvantages in physics or chemistry in the IST group, and there was even a newly established yet weaker disadvantage in biology. Thus, our study showed that IST does not generally reduce gender differences in self-concepts and interests in science.

We further hypothesized that IST might be a promising solution for closing the STEM gender gap and for strengthening female students' self-concepts and interests if "contagion effects" from biology to the other two domains can be achieved. However, our results indicate that this mechanism, if at play at all, may have worked in the opposite direction. That is, the presence of chemistry and physics may have reduced the attractiveness of biology for girls in the IST group. Previous studies have shown that many students have an image of physics as a "masculine" subject and that the gender role of girls in many cases is at odds with showing high achievement in physics because students tend to ask whether their own self-concept lines up with the prototype of students who are good at physics (Hannover & Kessels, 2004; Kessels, 2005; Kessels & Hannover, 2008). It might be the case that students believe integrated science is a more masculine subject as well, and this belief may have reduced female students' self-concepts and interest. However, future studies on the perception of integrated science would be required to test this hypothesis.

6.3 Limitations and Directions for Future Research

Our study was based on cross-sectional data, and thus, we cannot make causal claims based on these data. However, we cannot think of possible confounding variables that are related to course composition (which is mostly defined by the federal state) and the structure of academic self-concepts and interests. Furthermore, we replicated all analyses with balanced samples (see Supplement F). Still a longitudinal study of the effect of course composition would be desirable. This would allow researchers not only to make better causal claims but also to study the process of construct differentiation across different grades in more detail. For example, changes in the structure of self-concepts and interests when new science subjects are introduced (as mentioned above, in the SST group, biology is usually introduced first with chemistry and physics following in later grades) could be studied in

more detail. Furthermore, it could be studied whether the composition of science courses affects not only self-concepts and interests but also future (gendered) educational choices.

In addition, even though our study used a large representative data set, the subsample of students matching our strict criteria for the IST group was smaller than the SST group. As our additional analyses with other course composition groups showed, there appeared to be a qualitative difference between the strictly defined IST group and the groups that received IST in addition to SST (see Supplement E). Still, future studies should (over)sample this group to obtain a larger sample size.

Both in the Introduction and when discussing the theoretical implications, we proposed several mechanisms that might have contributed to the development of the differential structures of self-concept and interest. Unfortunately, owing to the cross-sectional approach and missing information on the specific practices across the science teaching groups, we could not test these mechanisms, which is a major limitation of our study. To fill this gap, future studies could, for example, apply a longitudinal design to determine whether the differentiation of constructs between the two groups slowly becomes stronger over time (i.e., test whether there are similar structures in the IST and SST groups in early secondary school that then become more distinct) or whether there are specific timepoints at which these differences are established (e.g., at the beginning of secondary school if the composition of students in a course already differs at this point or when new school subjects are introduced). Furthermore, researchers could examine the developmental interplay between motivational constructs: For example, is self-concept affected first, and are the effects then transferred to interest? Or are both constructs simultaneously affected by course composition? To test an additional mechanism, students across the groups could be asked how similar they consider the different domains of science to be. Pointing out similarities is a goal of IST, and perceived similarity has been shown to affect the extent to

which domain-specific self-concept is associated across different domains (Helm, Mueller-Kalthoff, Nagy, & Möller, 2016). In addition, this study could not explore how salient the domains “biology”, “chemistry” and “physics” were for students in the IST group given they don’t attend separated school subjects. Based on a comparison of topics covered in integrated science in early secondary school (Grasser, 2010), we know that the there is a large variability from topics that can clearly be assigned to one domain only (e.g., “plants and animals to biology”), to others topics such as “the sun” or “water” that are clearly interdisciplinary. In future studies, IST students could be interviewed how they would characterize the domains of biology, chemistry and physics. Think-aloud protocols could also clarify how students respond to domain-specific self-concept and interest items and what information about the domains they use.

From the perspective of teaching, future studies could explore whether teachers who teach integrated science classes differ from single-subject teachers in their didactic approaches or classroom management. Regarding the effects of merely labeling a course or subject as well as the effects of grading, experimental vignette studies might offer a viable approach. From the perspective of science education, it would be interesting to compare the typical curricula (which vary by federal state and school track) in the IST and SST approaches with regard to the topics covered, the contexts, and the activities students engage in.

7 References

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–816. https://doi.org/10.1037/a0021075

Arens, A. K., & Jansen, M. (2016). Self-concepts in reading, writing, listening, and speaking: A multidimensional and hierarchical structure and its generalizability across native and foreign languages. *Journal of Educational Psychology*, *108*(5), 646–664. https://doi.org/10.1037/edu0000081

Arens, A. K., & Morin, A. J. S. (2016). Examination of the Structure and Grade-Related Differentiation of Multidimensional Self-Concept Instruments for Children Using ESEM. *The Journal of Experimental Education*, *84*(2), 330–355. https://doi.org/10.1080/00220973.2014.999187

Beier, M. E., & Ackerman, P. L. (2003). Determinants of health knowledge: An investigation of age, gender, abilities, personality, and interests. *Journal of Personality and Social Psychology*, *84*(2), 439–448. https://doi.org/10.1037/0022-3514.84.2.439

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology*, *102*(4), 964–981. https://doi.org/10.1037/a0019644

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://doi.org/10.1037/bul0000052

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking Effects Depend on Tracking Type: An International Comparison of Students' Mathematics Self-Concept.

American Educational Research Journal, *50*(5), 925–957. https://doi.org/10.3102/0002831213489843

Chowdhary, B., Liu, X., Yerrick, R., Smith, E., & Grant, B. (2014). Examining Science Teachers' Development of Interdisciplinary Science Inquiry Pedagogical Knowledge and Practices. *Journal of Science Teacher Education*, *25*(8), 865–884. https://doi.org/10.1007/s10972-014-9405-0

Czerniak, C. M., & Johnson, C. C. (2011). Interdisciplinary Science Teaching. In *Handbook of Research on Science Education, Volume II*. Routledge. https://doi.org/10.4324/9780203097267.ch20

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I Like to Do It, I'm Able, and I Know I Am: Longitudinal Couplings Between Domain-Specific Achievement, Self-Concept, and Interest. *Child Development*, *78*(2), 430–447. https://doi.org/10.1111/j.1467-8624.2007.01007.x

Eccles, J. (2011). Gendered educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *International Journal of Behavioral Development*, *35*(3), 195–201. https://doi.org/10.1177/0165025411398185

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and Gender Differences in Children's Self- and Task Perceptions during Elementary School. *Child Development*, *64*(3), 830–847. https://doi.org/10.1111/j.1467-8624.1993.tb02946.x

Ehm, J.-H., Lindberg, S., & Hasselhorn, M. (2014). Reading, writing, and math self-concept in elementary school children: influence of dimensional comparison processes. *European Journal of Psychology of Education*, *29*(2), 277–294. https://doi.org/10.1007/s10212-013-0198-x

Elster, D. (2007). Student interests — the German and Austrian ROSE survey. *Journal of Biological Education*, *42*(1), 5–10. https://doi.org/10.1080/00219266.2007.9656100

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

Executive Agency Education, Audiovisual and Culture (Ed.). (2011). *Science education in Europe: national policies, practices and research*. Brussels: Eurydice [u.a.].

Grasser, A. (2010). Integrierte Naturwissenschaft: Entwicklung, Erprobung und Evaluation eines Projektunterrichts. Retrieved from https://www.db-thueringen.de/receive/dbt_mods_00017181

Guo, J., Marsh, H. W., Parker, P. D., Morin, A. J. S., & Dicke, T. (2017). Extending expectancy-value theory predictions of achievement and aspirations in science: Dimensional comparison processes and expectancy-by-value interactions. *Learning and Instruction*, *49*, 81–91. https://doi.org/10.1016/j.learninstruc.2016.12.007

Hannover, B., & Kessels, U. (2004). Self-to-prototype matching as a strategy for making academic choices. Why high school students do not like math and science. *Learning and Instruction*, *14*(1), 51–67. https://doi.org/10.1016/j.learninstruc.2003.10.002

Hardy, G. (2014). Academic Self-Concept: Modeling and Measuring for Science. *Research in Science Education*, *44*(4), 549–579. https://doi.org/10.1007/s11165-013-9393-7

Häussler, P., & Hoffmann, L. (2000). A curricular frame for physics education: Development, comparison with students' interests, and impact on students' achievement and self-concept. *Science Education*, *84*(6), 689–705. https://doi.org/10.1002/1098-237X(200011)84:6<689::AID-SCE1>3.0.CO;2-L

Helm, F., Mueller-Kalthoff, H., Nagy, N., & Möller, J. (2016). Dimensional Comparison Theory: Perceived Subject Similarity Impacts on Students' Self-Concepts. *AERA Open*, *2*(2), 2332858416650624. https://doi.org/10.1177/2332858416650624

Hidi, S., & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

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*(5888), 494–495. https://doi.org/10.1126/science.1160364

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

Jansen, M., Lüdtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology*, *46*, 116–127. https://doi.org/10.1016/j.cedpsych.2016.05.004

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

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2015). Contrast and assimilation effects of dimensional comparisons in five subjects: An extension of the I/E model. *Journal of Educational Psychology*, *107*(4), 1086–1101. https://doi.org/10.1037/edu0000021

Kessels, U. (2005). Fitting into the stereotype: How gender-stereotyped perceptions of prototypic peers relate to liking for school subjects. *European Journal of Psychology of Education*, *20*(3), 309–323. https://doi.org/10.1007/BF03173559

Kessels, U., & Hannover, B. (2008). When being a girl matters less: Accessibility of gender-related self-knowledge in single-sex and coeducational classes and its impact on students' physics-related self-concept of ability. *British Journal of Educational Psychology*, *78*(2), 273–289. https://doi.org/10.1348/000709907X215938

Krapp, A. (2002). Structural and dynamic aspects of interest development: theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, *12*(4), 383–409. https://doi.org/10.1016/S0959-4752(01)00011-1

Krapp, A., & Prenzel, M. (2011). Research on Interest in Science: Theories, methods, and findings. *International Journal of Science Education*, *33*(1), 27–50. https://doi.org/10.1080/09500693.2010.518645

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Lazarides, R., & Ittel, A. (2012). Instructional Quality and Attitudes toward Mathematics: Do Self-Concept and Interest Differ across Students' Patterns of Perceived Instructional Quality in Mathematics Classrooms? *Child Development Research*, *2012*, e813920. https://doi.org/10.1155/2012/813920

Lenski, A. E., Matrin Hecht, Penk, C., Milles, F., Mezger, M., Heitmann, P., … Pant, H. A. (2016). IQB-Ländervergleich 2012 – Skalenhandbuch zur Dokumentation der Erhebungsinstrumente. Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.20386/HUB-42547

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, *82*(4), 623–636. https://doi.org/10.1037/0022-0663.82.4.623

Marsh, H. W., & Ayotte, V. (2003). Do Multiple Dimensions of Self-Concept Become More Differentiated With Age? The Differential Distinctiveness Hypothesis. *Journal of Educational Psychology*, *95*(4), 687–706. https://doi.org/10.1037/0022-0663.95.4.687

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, *54*(2), 263–280. https://doi.org/10.1037/dev0000393

Möller, J., & Köller, O. (2001). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. *Journal of Educational Psychology*, *93*(4), 826–835. https://doi.org/10.1037//0022-0663.93.4.826

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

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the “x” out of expectancy-value theory?: A psychological mystery, a

substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22*(8), 1058–1066. https://doi.org/10.1177/0956797611415540

OECD. (2015). *The ABC of Gender Equality in Education*. OECD Publishing. https://doi.org/10.1787/9789264229945-en

Opdenakker, M.-C., Maulana, R., & Brok, P. den. (2012). Teacher–student interpersonal relationships and academic motivation within one school year: developmental changes and linkage. *School Effectiveness and School Improvement*, *23*(1), 95–119. https://doi.org/10.1080/09243453.2011.619198

Pant, H. A., Stanat, P., Hecht, M., Heitmann, P., Jansen, M., Lenski, A. E., … Siegle, T. (2015). *IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 (IQB-LV 2012).* Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_LV_2012_v1.

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

Ramm, G. C., Adamsen, C., Neubrand, M., & Deutsches PISA-Konsortium (Eds.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept – Testing the reciprocal effects model. *Learning and Instruction*, *29*, 21–30. https://doi.org/10.1016/j.learninstruc.2013.07.004

Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, *13*(4), 425–438. https://doi.org/10.1080/15248372.2012.717340

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Schmidt, I., Brunner, M., Keller, L., Scherrer, V., Wollschläger, R., Baudson, T. G., & Preckel, F. (2017). Profile formation of academic self-concept in elementary school students in grades 1 to 4. *PLOS ONE*, *12*(5), e0177854. https://doi.org/10.1371/journal.pone.0177854

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

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

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

Stipek, D., & Mac Iver, D. (1989). Developmental Change in Children's Assessment of Intellectual Competence. *Child Development*, *60*(3), 521–538. https://doi.org/10.2307/1130719

Strickhouser, J. E., & Zell, E. (2015). Self-evaluative effects of dimensional and social comparison. *Journal of Experimental Social Psychology*, *59*, 60–66. https://doi.org/10.1016/j.jesp.2015.03.001

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology*, *104*(3), 763–777. https://doi.org/10.1037/a0027470

Tsai, Y.-M., Kunter, M., Lüdtke, O., Trautwein, U., & Ryan, R. M. (2008). What makes lessons interesting? The role of situational and individual factors in three school subjects. *Journal of Educational Psychology*, *100*(2), 460–472. https://doi.org/10.1037/0022-0663.100.2.460

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. https://doi.org/10.1016/j.dr.2013.08.001

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*, 770–775. https://doi.org/10.1177/0956797612458937

Wigfield, A., & Eccles, J. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wolff, F., Helm, F., Zimmermann, F., Nagy, G., & Möller, J. (2018). On the Effects of Social, Temporal, and Dimensional Comparisons on Academic Self-Concept. *Journal of Educational Psychology*, 1005–1025. https://doi.org/10.1037/edu0000248

8 Tables

Table 1

Fit Indices of Measurement Invariance Tests of Academic Self-Concept and Interest in Science Across the IST and the SST Groups

	Model	S-B χ^2	*df*	Δ S-B χ^{2a}	CFI	RMSEA	SRMR
1	Configural invariance model (six first-order factors; self-concept and interest in biology, chemistry, and physics)	1,325.22*	426	-	.983	.030	.028
2	Weak invariance model	1,362.06*	444	32.18	.982	.030	.030
3	Strong invariance model	1,402.53*	462	39.52*	.982	.029	.030
4	Strict invariance model	1,438.70*	486	41.98	.982	.029	.030
5	Strict invariance model with equal factor variances	1,448.85*	492	6.92	.982	.029	.032
6	Strict invariance model with equal factor variances and covariances	1,639.61*	507	136.41*	.978	.031	.053

Note. S-B χ^2 = Satorra-Bentler χ^2; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. n_{SST} = 4,361; n_{IST} = 337

[a]Comparison with model above.

* $p < .01$.

Table 2

Factor Correlations for Students From the Separate Teaching Group (Below the Diagonal) and the Integrated Teaching Group (Above the Diagonal)

	1	2	3	4	5	6
1. Academic self-concept in biology		.75*	.71*	.67*	.53*	.48*
2. Academic self-concept in chemistry	.35*		.82*	.52*	.71*	.53*
3. Academic self-concept in physics	.30*	.45*		.48*	.57*	.61*
4. Interest in biology	.69*	.25*	.19*		.66*	.68*
5. Interest in chemistry	.26*	.77*	.39*	.37*		.87*
6. Interest in physics	.19*	.35*	.76*	.30*	.54*	

Note. The correlations are from a latent Confirmatory Factor Analysis model with strict measurement invariance and equal factor variances (Table 1, Model 5). n_{SST} = 4,361; n_{IST} = 337

*$p < .01$

.

9 Figures

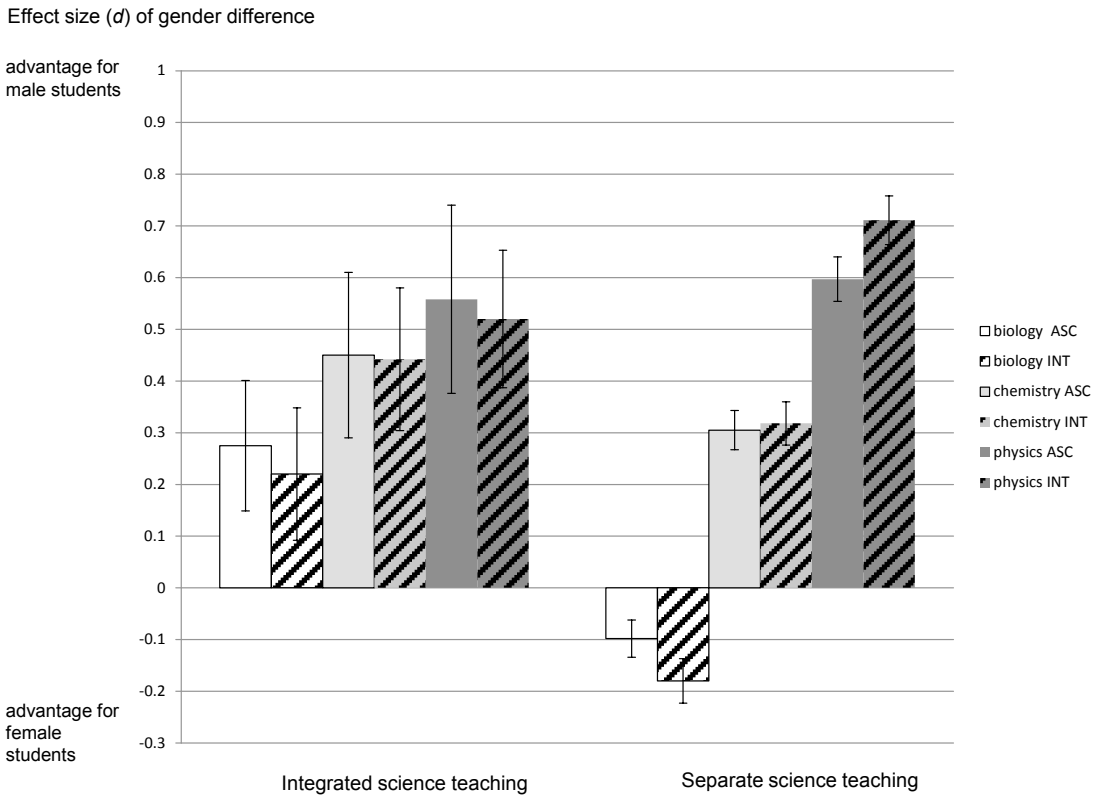

Figure 1. Effect sizes (Cohen's *d*) of gender differences in latent means for biology, chemistry, and physics self-concept (ASC) and interest (INT). Positive values indicate higher academic self-concept for male students, whereas negative values indicate higher academic self-concept for female students. Errors bar represent standard errors. All estimates are based on a strong measurement invariance model (Table A1, Model 3). $n_{\text{SST-male}}$ = 1,390; $n_{\text{SST-female}}$ = 1,254, $n_{\text{SST-male}}$ = 101; $n_{\text{SST-female}}$ = 76.

The Dimensional Structure of Students' Self-Concept and Interest in Science Depends on Course Composition

Online Supplement

Supplement A: The German secondary school system

After primary school, German students enter a tracked secondary school system. Because the German federal states (rather than the national government) are in charge of regulating the school systems, there are large differences between the federal states with regard to school tracks and types. For example, in some states, students enter the secondary school system after 4 years of primary school, whereas in other states, 6 years of primary school are mandatory. In the first grade of primary school, the majority of students are 6 years old (the cut-off dates vary slightly by state; students who have turned 6 before the cut-off date are enrolled in first grade at the next available point in time). Thus, they move on to the secondary school system on average at age 10 or 12 depending on whether there are 4 or 6 years of primary school in a given federal state. Some states differentiate between only the academic track (*Gymnasium*) and one other secondary school type (e.g., *Sekundarschule*) in which within-school tracking is employed, whereas other states employ explicit between-school tracking with several other secondary school types in addition to *Gymnasium* (e.g., *Realschule*, *Werkrealschule*, *Hauptschule*). Without considering all of the different school types and complexities of the German school system that are explained and discussed elsewhere (Maaz, Trautwein, Lüdtke, & Baumert, 2008), the differentiation between Gymnasium as the primary academic, university-bound track and the other more vocationally oriented school types (also referred to as the nonacademic track) is present in all federal states and is a central characteristic of the German school system.

Science is taught in both school types such that the total number of lessons is, on average, slightly higher in the academic track. The practice of teaching integrated science

exists only in the nonacademic track schools, which is why we restricted our sample to these schools (see Section 4.3). In both school tracks, separated science teaching (SST) and mixed forms (also see Supplement E) are more prevalent than the strict form of integrated science teaching (IST) where biology, chemistry, and physics are not taught in addition to the integrated subject.

However, it should be noted that the national educational standards for the intermediate school exam after grade 10 which form the core of the curriculum are valid for both school types. Furthermore, previous research showed no differences in the structure of academic self-concept in science between these two school types (Jansen, Schroeders, & Lüdtke, 2014).

Supplement B: Procedure for Measurement Invariance

Invariance of factor correlations. To answer our first research questions concerning a possibly less differentiated factor structure in the IST group compared to the SST group, we ultimately aimed to test the factor correlations for measurement invariance. In order to compare such parameters of the structural model we were interested in, we first tested for invariance of the measurement parameters across the two groups following recommendations from the literature (Little, 2013; van de Schoot, Lugtig, & Hox 2012). We started with a configural invariance model in which the same factor structure was assumed in both groups, but all parameters of the measurement model were freely estimated. We then constrained the factor loadings (weak/metric invariance model), intercepts (strong/scalar invariance model), and residual variances (strict invariance model) to be equal across the two groups and compared each model's fit with the less restrictive model. In a further step, we tested for equality of variances across groups and, most important, we tested for equality of covariances across groups. Given an invariance of variances, these differences in covariances could then be interpreted as differences in correlations. The results are shown in Table 1 of the manuscript.

Invariance of Gender Differences in Means. To answer our second major research question about whether there would be differential gender differences between the two groups, we split the sample into a total of four groups (IST/female, IST/male, SST/female, SST/male). Before comparing the self-concept and interest means across groups, we first tested for measurement invariance across the four groups because the invariance of intercepts (strong invariance) is a prerequisite for meaningful mean comparisons (van de Schoot et al., 2012). The results are described in Supplement D. We then compared the effect size of the gender difference (Cohen's *d* adapted for latent mean comparisons, Hancock, 2001) across

the IST and SST groups for all three domains using z-tests. The results are described in detail in the result section of the manuscript.

Model evaluation. To evaluate the absolute goodness of fit to the data for each of these models, we examined the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Square Residual (SRMR) using the guidelines described by Hu and Bentler (1999). They propose that a CFI above .95, an RMSEA below .06, and an SRMR below .08 indicate that a model has good fit. To compare nested models, we examined different criteria: the Satorra-Bentler (S-B) scaled χ^2-difference test, changes in the CFI and the RMSEA, as well as absolute changes in the model parameters (e.g., factor correlations). We weighted these criteria differently according to the models we compared. Because the χ^2-difference test tends to be very sensitive even to trivial model misfit when many model parameters are manipulated at once as in invariance testing, we followed Little's (2013) recommendation to examine changes in the CFI and the RMSEA when evaluating the invariance of measurement parameters (factor loadings, intercepts, residual variances). The suggestions for differences that indicate that the null hypothesis of model equality should not be rejected range from $\Delta CFI < .01$ (Cheung & Rensvold, 2002) to $\Delta CFI < .002$ (Meade, Johnson, & Braddy, 2008) for the CFI. For the RMSEA, there is only one recommendation by Cheung and Rensvold (2002), who suggested that a $\Delta RMSEA$ of .015 or less represents a nonsignificant decrease in model fit. However, because these two indices might not show enough sensitivity to detect differences when testing for invariance in structural parameters such as means and factor correlations (when only a few parameters are manipulated at once), we used the χ^2-difference test as recommended by Little (2013). We also examined the absolute parameter values and calculated effect sizes to assess whether the parameter differences seemed relevant and meaningful.

Supplement C: Descriptive Statistics

The means, standard deviations, and reliability coefficients can be found in Table C.1. As mentioned in the method section, all scales showed high reliability with ω values of .85 and higher in both groups. Thus, there seemed to be no reliability differences across groups.

With scale means ranging from 2.24 to 2.78 (on the 4-point scale), there were also no ceiling or floor effects. Both self-concept and interest were highest in biology and lowest in physics. There were only slight mean differences between the two groups which showed no consistent main effect of higher self-concepts or interests in any of the two groups. However, for both self-concept and interest, the means were less differentiated between domains in the IST group (e.g., the difference between the self-concept means in biology and physics was smaller in the IST group compared with the SST group) already hinting at possible differences in the factorial structure.

In addition, there were no differences in gender or SES (operationalized by the HISEI index; Ganzeboom, De Graaf, & Treiman, 1992), but students from the IST group were more likely to have an immigrant background than students from the SST group. This is likely due to the combination of school type and federal state in which IST is implemented. We have no hypothesis why this difference should theoretically affect students' self-concept and interest structures, but replicated all analyses using Propensity Score Matching to test this assumption and found all results to be replicable in the matched samples (see Supplement F)

Table C.1

Descriptive Statistics for Students From the Separate Teaching Group (SST) and the Integrated Teaching Group (IST)

	SST group					IST group				
	M	*SD*	ω	%	N	*M*	*SD*	ω	%	N
Motivational constructs										
Academic self-concept in biology	2.78	0.67	.86		4,282	2.68	0.65	.85		329
Academic self-concept in chemistry	2.55	0.77	.88		4,315	2.58	0.71	.86		327
Academic self-concept in physics	2.52	0.74	.87		4,332	2.57	0.70	.86		329
Interest in biology	2.51	0.78	.91		2,572	2.35	0.84	.93		173
Interest in chemistry	2.29	0.83	.92		2,607	2.34	0.90	.94		171
Interest in physics	2.24	0.83	.93		2,614	2.30	0.85	.94		170
Demographics										
Gender female				47%	4,361				45%	337
Immigrant background (at least one parent born outside of Germany)				23%	3,813				37%	262
HISEI	42.88	18.48			3,568	45.56	19.05			257

Note. ω = McDonald's omega as a reliability measure (McDonald, 1999). N = sample sizes with valid values for given variable. Mean scores, standard deviations and sample sizes for self-concept and interest scales refer to manifest scores using listwise deletion.

Supplement D: Results for Measurement Invariance Testing Across Gender and Course Composition Groups

Following the procedure described in Supplement B, we tested for measurement invariance across the four groups (gender crossed with course composition) by introducing a series of increasingly stricter assumptions and evaluating these models with regard to our criteria (i.e., differences in CFI and the RMSEA for the measurement parameters, see Supplement B). The model fit results are shown in Table D.1. At least strong measurement invariance is recommended for a valid interpretation of mean differences and we were indeed able to establish this level of invariance (configural vs. strong invariance: ΔCFI = -.006, ΔRMSEA = -.002, see Table D.1).

Table D.1
Measurement Invariance Testing for Four-Group Models (Course Composition x Gender)

	Model	S-B χ^2	*df*	Δ S-B χ^{2a}	CFI	RMSEA	SRMR
1	Configural invariance model across all four groups (six first-order factors, self-concept and interest in biology, chemistry, and physics)	1,938.91*	852	-	.979	.033	.032
2	Weak invariance model all four groups	2,149.59[a]*	906	215.11*	.976	.034	.038
3	Strong invariance model all four groups	2,364.11[a]*	960	439.72*	.973	.035	.039

Note. S-B χ^2 = Satorra-Bentler χ^2, CFI = comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.
[a]Comparison with model above.
* $p < .01$.

Supplement E: Robustness Check with Additional Course Composition Groups

As described in detail in the methods section (more precisely, section 4.3), our study focused on the comparisons of two strictly, but clearly defined groups. Students who were taught science exclusively as an integrated subject in all secondary school grades (Integrated Science Teaching group, IST, N = 337) and students who were taught science exclusively through separate lessons in biology, chemistry and physics across all secondary school grades without ever attending an integrated science class (Separate Science Teaching group, SST, N = 4,361). Even though we think this comparison is the easiest and clearest to interpret, we identified two further groups that can be examined amongst the 2,934 students that could not be classified as strictly IST or SST (see section 4.3 of the manuscript). From this group, 559 had received IST in grades 5 and 6 following by SST in grades 7 to 9. This approach to begin with integrated science classes and then move on to teaching separate subjects is quite common in Germany. For the following comparisons, we will refer to this group as "IST-SST". The remaining 2,375 students had received other, less systematic combinations of IST and SST. This includes, for example, an integrated science class (e.g., focusing on how science contributes to solving real-work problems) offered in addition to the more traditional SST approach. We thus termed this group as "Other" in the following.

In our study, we made clear predictions about possible differences in the factor structures between the IST and the SST group with a more differentiated structure in the SST group and a less differentiated structure in the IST groups (showing in higher factor relations). We find strong evidence for this assumption. However, what would we expect for the remaining groups? Given that they are somewhat "in between" with regard to the exposure to IST that they had, we could expect the factor correlations in these groups to be weaker than in the IST group, but stronger than in the SST group. This would imply, however, that there are lasting effects of IST on the factor structure even if it hasn't been taught for several school years or

is only taught in addition to SST. Table E.1 shows the latent factor correlations from a multi-group model involving the four above mentioned groups.

For the SST and the IST group, the results from the 4-group model replicate the results reported in the manuscript which is to be expected (see Table E.1). That is, the correlations in the IST group were clearly higher compared to the SST group indicating a less differentiated factor structure for both self-concept and interest.

In the other two groups, overall, the results were quite similar to the SST group. That is, the correlations between self-concept and interest measures in the three science domains were mostly below .50 (with the correlation between interest in physics and chemistry in the "Other" group being an exception with .61 which is, however, still clearly lower than the .87 from the IST group). While some values were numerically higher in the "Other" group compared to the SST and IST-SST groups (and the other way around), the differences were not systematic whereas the pattern of higher correlations in the IST group was clearly present.

When fixing the factor correlations in the SST, the IST-SST and the "Other" group to equality and only estimating the factor correlations in the IST group separately, the model fit, assuming strict invariance of the measurement parameters, did not decrease significantly compared to a model in which the correlations in all groups were estimated freely ($\Delta\chi^2$ = 47.62, df = 30, p = .02). This omnibus test shows that there are, overall, no significant differences in correlations between these three groups and the results are essentially similar to the SST group. When we also fix the correlations in the IST group to equality with the other three groups (i.e., all correlations being equal across groups), the model fit clearly decreases ($\Delta\chi^2$ = 132.56 df = 15, p < .01) emphasizing the different factor structure in the IST group.

Table E.1

Factor Correlations for Students for Four Groups

	1	2	3	4	5
SST					
1. Academic self-concept in biology					
2. Academic self-concept in chemistry	.35*				
3. Academic self-concept in physics	.30*	.45*			
4. Interest in biology	.69*	.25*	.19*		
5. Interest in chemistry	.26*	.77*	.39*	.37*	
6. Interest in physics	.19*	.35*	.76*	.30*	.54*
IST					
2. Academic self-concept in chemistry	.75*				
3. Academic self-concept in physics	.71*	.82*			
4. Interest in biology	.67*	.52*	.48*		
5. Interest in chemistry	.53*	.71*	.57*	.66*	
6. Interest in physics	.48*	.53*	.61*	.68*	.87*
IST-SST					
2. Academic self-concept in chemistry	.33*				
3. Academic self-concept in physics	.42*	.40*			
4. Interest in biology	.65*	.25*	.23*		
5. Interest in chemistry	.18*	.80*	.22*	.30*	
6. Interest in physics	.24*	.33*	.71*	.24*	.43*

Other					
2. Academic self-concept in chemistry	.42*				
3. Academic self-concept in physics	.31*	.51*			
4. Interest in biology	.73*	.29*	.22*		
5. Interest in chemistry	.33*	.78*	.46*	.40*	
6. Interest in physics	.26*	.45*	.81*	.32*	.61*

Note. The correlations are from a latent Confirmatory Factor Analysis model with strict measurement invariance and equal factor variances. SST = separate science teaching in all grades, IST = integrated science teaching in all grades, IST-SST = integrated science teaching in grades 5 and 6 following by separate science teaching in grades 7 to 9, Other = not assignable to any of the three above mentioned groups due to a different combination of integrated and separate science teaching

*$p < .01$.

Supplement F: Robustness Check with Matched Samples

Possible pre-selection group differences are always a threat to causal interpretation in non-experimental studies. Since attending IST and SST science classes is typically not a choice, but depending on federal state or school, and since our outcome is not the mean but the factorial structure of academic self-concepts and interests, we do not think there is a strong argument for confounding variables influencing our results (as that would have to be variables on federal state level or that are somehow confounded with the science teaching method that would reasonably affect the structure of these motivational structures).

However, we still aimed to check this assumption empirically using one recommended method for dealing with possibly confounding group differences in non-experimental settings—propensity score matching (PSM). The goal of PSM is to balance out preexisting differences between a treatment group (in our case, the IST students) and a control group (in our case, the SST students) by selecting and/or weighting cases from the control group in a specific way to ensure a high comparability to the treatment group (finding "statistical twins"). An advantage of this procedure is that it is used as a non-parametric data augmentation method before the data analysis itself. Therefore, it can include a larger set of control variables, possibly some of them being highly related, without adding further complexity to the analysis model itself (which is only run on the matched sample afterwards) or causing multicollinearity.

There are a variety of different PSM methods and algorithms that are described elsewhere in detail (e. g., Austin, 2011, Stuart, 2010). We decided for one of the most common approaches that has the advantage of using all cases from the treatment (which is quite small in our case) and the control group and balances the groups by a weighting procedure—full matching (see Stuart & Green, 2008) using the R-package MatchIt (Ho, Imai, King, & Stuart, 2011).

Since we did not have any specific hypotheses as mentioned above, we aimed to use a broad mix of covariates relating both to the students' school performance and their background characteristics. More precisely, we included in the matching model gender, achievement in four different school subjects, an index of home possessions (HOMEPOS as used in the PISA studies, described in detail in the PISA 2012 technical report), migration background, number of books at home, language other than German spoken at home (0 = never, 1 = sometimes/ofen/always), the highest ISEI index (Ganzeboom et al., 1992) and EGP class (Erikson, Goldthorpe, & Portocarero, 1979) of the parents and the score in a short test of cognitive ability (a figural analogies test). For an overview, also see Figure F.1.

First, we checked for differences across the groups on these covariates before and after matching. An established procedure for this check is comparing standardized mean differences (sometimes also referred to as standardized bias). As Figure F.1 shows, the matching procedure was successful in reducing the mean differences between the two groups substantially.

We then replicated the two central analyses of the paper—the comparison of factor correlations (see Table F.1) and the comparison of gender differences in factor means (see Figures F.2)—with the matched samples (using the WEIGHTS option in Mplus). The result pattern was very similar to the unmatched samples even though there were some numerical differences in correlations the SST group (the results in the IST group were identical as it was treated at the treatment group and differential weighting through the matching procedure was thus only employed for the SST group).

Table F.1

Factor Correlations for Students From the Separate Teaching Group (Below the Diagonal) and the Integrated Teaching Group (Above the Diagonal) Using Matched Samples (Matching Weights from a Full Matching Model)

	1	2	3	4	5	6
1. Academic self-concept in biology		.75*	.70*	.67*	.53*	.46*
2. Academic self-concept in chemistry	.27*		.80*	.50*	.71*	.49*
3. Academic self-concept in physics	.26*	.38*		.46*	.56*	.57*
4. Interest in biology	.76*	.26*	.24*		.65*	.66*
5. Interest in chemistry	.13*	.79*	.39*	.27*		.85*
6. Interest in physics	.15*	.26*	.73*	.23*	.52*	

Note. The correlations are from a latent Confirmatory Factor Analysis model with strict measurement invariance and equal factor variances

*$p < .01$.

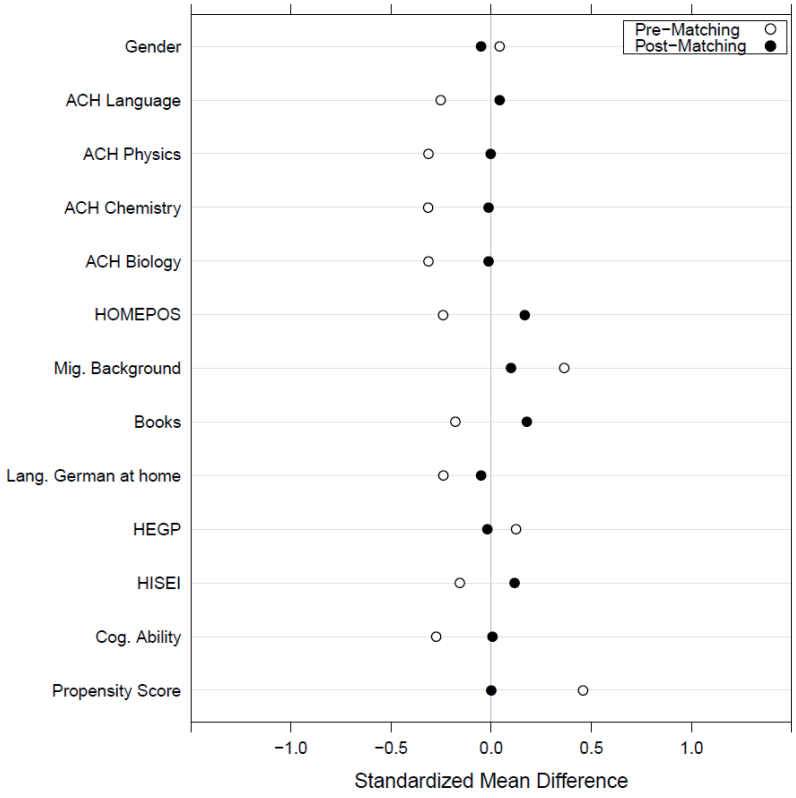

Figure F.1. Standardized mean differences in covariates and in the propensity score before (white) and after (black) the full matching procedure. Positive values indicate higher means in the treatment group (IST); negative values indicate higher means in the control group (SST). ACH = Achievement (achievement tests), HOMEPOS = home possession index from PISA 2012, Mig. Background = migration background (at least one parent born outside of Germany, 0 = no, 1 = yes), HEGP = higher EGP class (Ericson-Goldthorpe- Portocarero scheme for classifying social classes) of the two parents, HISEI = higher International Socio-economic Index of Occupational Status of the two parents, cog. ability = results from a figural analogies test as proxy of general cognitive ability.

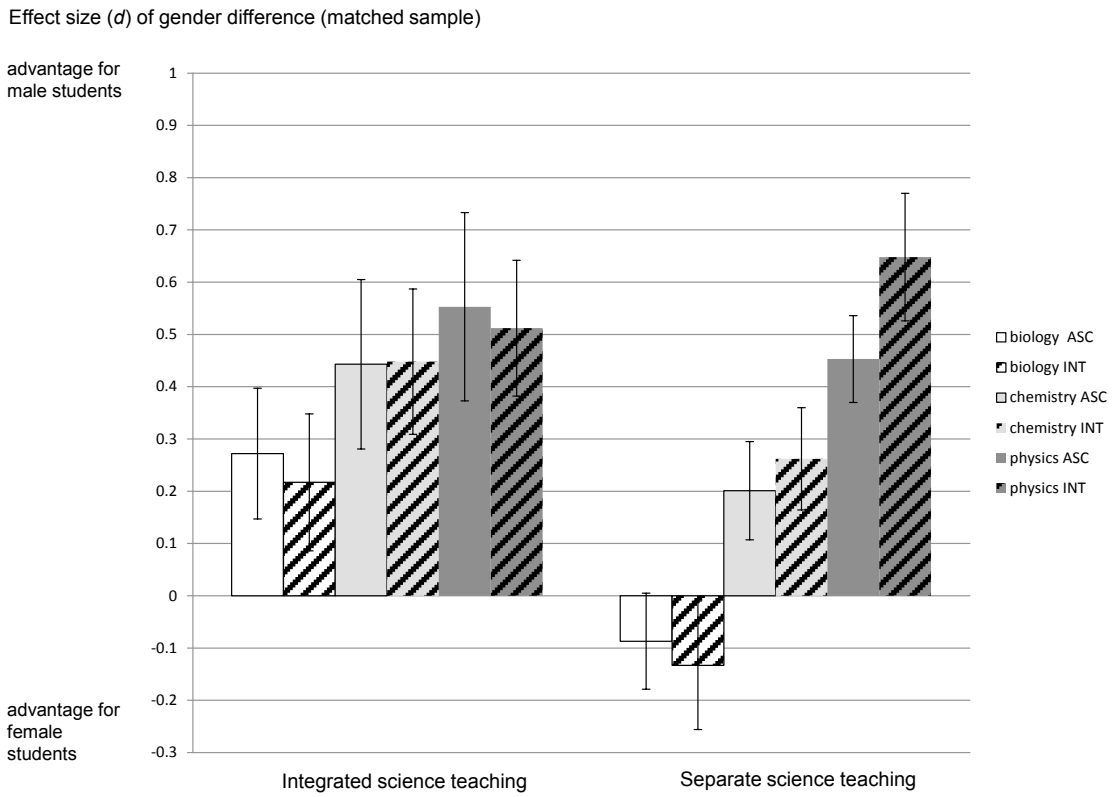

Figure F.2. Effect sizes (Cohen's *d*) of gender differences in latent means for biology, chemistry, and physics self-concept (ASC) and interest (INT) for the matched samples. Positive values indicate higher academic self-concept for male students, whereas negative values indicate higher academic self-concept for female students. Errors bar represent standard errors. All estimates are based on a strong measurement invariance model.

Supplement G: Sensitivity Analyses for Missing Data Treatment

As mentioned in detail in the manuscript, a planned missing data design was used which included several questionnaire versions. Self-concept was included in three of the eight versions and interest was included in two.

Because we were aiming to use a Full Information Maximum Likelihood (FIML) estimation to treat the missing data while retaining as much power as possible, we did not exclude any students as long as they had a valid value for one of self-concept items. To check whether this selection of the base sample (from which the IST and SST samples were then selected), had any effect, we compared the main results—that is, the pattern of factor correlations—for this sample to three further approaches: (a) using only students who had valid values on at least one self-concept and one interest item, (b) separating two separate three-factor models in the self-concept and the interest sample (each defined by at least one valid item) and (c) a sample in which listwise deletion was employed (i.e., we analyzed only students who had valid values on all items). The results can be found in Table G.1. The pattern of correlations was very similar across all approaches suggesting a high robustness of the results.

Table G.1

Factor Correlations for Students From the Separate Teaching Group (Below the Diagonal) and the Integrated Teaching Group (Above the Diagonal) in Different Samples

	1	2	3	4	5	6
Base sample: At least one self-concept item (FIML) N_{SST} = 4,361; N_{IST} = 337						
1. Academic self-concept in biology		.75*	.71*	.67*	.53*	.48*
2. Academic self-concept in chemistry	.35*		.82*	.52*	.71*	.53*
3. Academic self-concept in physics	.30*	.45*		.48*	.57*	.61*
4. Interest in biology	.69*	.25*	.19*		.66*	.68*
5. Interest in chemistry	.26*	.77*	.39*	.37*		.87*
6. Interest in physics	.19*	.35*	.76*	.30*	.54*	
At least one self-concept AND one interest item (FIML); 6-factor model N_{SST} = 2,644; N_{IST} = 177						
1. Academic self-concept in biology		.76*	.73*	.68*	.54*	.49*
2. Academic self-concept in chemistry	.36*		.84*	.53*	.71*	.54*
3. Academic self-concept in physics	.32*	.51*		.49*	.58*	.61*
4. Interest in biology	.69*	.26*	.19*		.67*	.68*
5. Interest in chemistry	.27*	.77*	.43*	.37*		.87*
6. Interest in physics	.21*	.40*	.77*	.31*	.57*	

At least one self-concept OR one interest item (FIML); two separate 3-factor models; **Self-concept: N_{SST} = 4361; N_{IST} = 337** **Interest: N_{SST} = 2,638; N_{IST} = 177**	1	2	3	4	5	6
1. Academic self-concept in biology		.76*	.71*			
2. Academic self-concept in chemistry	.35*		.82*			
3. Academic self-concept in physics	.30*	.45*				
4. Interest in biology					.67*	.68*
5. Interest in chemistry				.37*		.87*
6. Interest in physics				.31*	.57*	
Listwise deletion (i.e., no missing data on any self-concept or interest item); **N_{SST} = 2,300; N_{IST} = 156**						
1. Academic self-concept in biology		.76*	.72*	.67*	.53*	.47*
2. Academic self-concept in chemistry	.36*		.84*	.51*	.70*	.51*
3. Academic self-concept in physics	.32*	.51*		.48*	.57*	.60*
4. Interest in biology	.70*	.24*	.21*		.67*	.67*
5. Interest in chemistry	.26*	.78*	.43*	.37*		.86*
6. Interest in physics	.21*	.39*	.77*	.30*	.57*	

Note. The correlations are from a latent Confirmatory Factor Analysis model with strict measurement invariance and equal factor variances (analogous to Table 1, Model 5). SST = Separate science teaching group; IST = integrated science teaching group.

*$p < .01$

Supplement H: MIMIC Model to Study the Interaction Between Course Composition and Gender Effects

As an additional test of the robustness of the pattern of gender differences we found on the multigroup approach (see section 5.2 in the manuscript), we estimated a multiple indicator multiple cause (MIMIC) model. In this model, we used gender, course composition (IST vs. SST), and their interaction as predictors of the six self-concept and interest factors. The results (see Table H.1) showed main effects of gender (advantages for male students in chemistry and physics; advantages for female students in biology) but significant interaction effects only on self-concept and interest in biology but not in chemistry and physics, replicating the patterns found in the multigroup approach

Table H.1
Multiple Indicator Multiple Cause Models Predicting Self-Concept and Interest in Biology, Chemistry, and Physics

Predictors	B ASC		B INT		C ASC		C INT		P ASC		P INT	
	β	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Gender (0 = female, 1 = male)	-0.10	0.04	-0.18*	0.04	0.30*	0.04	0.31*	0.04	0.57*	0.04	0.67*	0.04
Course composition (0 = SST, 1 = IST)	-0.74*	0.21	-0.97*	0.21	-0.21	0.25	-0.34	0.26	0.08	0.27	0.19	0.20
Gender x Course Composition	0.37*	0.13	0.48*	0.13	0.15	0.14	0.23	0.14	-0.03	0.16	-0.07	0.12
R^2	.01		.01		.03		.03		.08		.11	

Note. B = biology, C = chemistry, P = physics, ASC = academic self-concept, INT = interest. Model fit indices of this model: S-B χ^2 = 1,106.03 (*df* = 267, $p < .01$), CFI = .984, RMSEA = .026, SRMR = .025. All coefficients are based on the STDY standardization from Mplus (Standardization of the dependent but not the independent variables). n = 4,698
* $p < .01$.

References appearing in the Supplement

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424. doi: 10.1080/00273171.2011.568786

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational Class Mobility in Three Western European Societies: England, France and Sweden. *The British Journal of Sociology*, *30*(4), 415. https://doi.org/10.2307/589632

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373–388. https://doi.org/10.1007/BF02294440

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*(8). https://doi.org/10.18637/jss.v042.i08

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. Doi: 10.1080/10705519909540118

Jansen, M., Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences.

Learning and Individual Differences, *30*, 11–21. https://doi.org/10.1016/j.lindif.2013.12.003

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: The Guilford Press.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational Transitions and Differential Learning Environments: How Explicit Between-School Tracking Contributes to Social Inequality in Educational Outcomes. *Child Development Perspectives*, *2*, 99–106. doi:10.1111/j.1750-8606.2008.00048.x

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah, N.J: L. Erlbaum Associates.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. Doi:10.1037/0021-9010.93.3.568

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1–21. Doi: 10.1214/09-STS313

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406. Doi: 10.1037/0012-1649.44.2.395

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi: 10.1080/17405629.2012.686740