European Journal of Science and Mathematics Education, 2024, **12**(1), 38-59
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Does a centralized written final examination in mathematics indeed improve pupils' subject-related study ability?

Pia Tscholl [1*]
0000-0002-0285-5059

Florian Stampfer [1]
0000-0003-4980-4272

Tobias Hell [2]
0000-0002-2841-3670

[1] Universität Innsbruck, Innsbruck, AUSTRIA
[2] DataLab Hell GmbH, Axams, AUSTRIA
[*] Corresponding author: pia.tscholl@uibk.ac.at

Citation: Tscholl, P., Stampfer, F., & Hell, T. (2024). Does a centralized written final examination in mathematics indeed improve pupils' subject-related study ability?. *European Journal of Science and Mathematics Education, 12*(1), 38-59. https://doi.org/10.30935/scimath/13829

ARTICLE INFO

Received: 19 May 2023
Accepted: 9 Oct 2023

ABSTRACT

Since 2015/16, a standardized written final examination in mathematics or applied mathematics has been compulsory for nearly all pupils at the upper secondary level in Austria. While this standardized competence-oriented maturity examination is intended to increase pupils' subject-related study ability, empirical research in this regard is scarce. Therefore, the subject-related study ability for six partially different control and experimental groups containing between 11 and 17 first-year STEM students is compared using a one-tailed two-sample Wilcoxon rank sum test. No significant differences in the subject-related study ability are detected between the control groups, comprising first-year Austrian STEM students who did not participate in the standardized written final examination in mathematics, and the experimental groups, comprising first-year Austrian STEM students who did participate in the standardized written final examination in mathematics. However, post hoc power analyses show that the sample sizes for each of the six sample cases would have to be much larger to prove significant results with a power of at least 80%. Additionally, no evidence for teaching-to-the test practices could be found in the experimental groups.

Keywords: standardized tests, mathematics, study skills, college readiness, STEM, teaching to the test

INTRODUCTION, RELEVANCE, & RESEARCH GAP

"Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen" (OECD, 2003, p. 24). To test basic mathematical knowledge and literacy among secondary school graduates according to this understanding, the *standardisierte kompetenzorientierte schriftliche Reifeprüfung bzw. Reife- und Diplomprüfung* [standardized competence-oriented written maturity examination, SRP] was designed and introduced in Austria (SRDP, 2021). The main result of this design process is a standardized written mathematics exam, which is based on a core competence catalog.

This standardized written mathematics exam serves as a compulsory final examination at *Allgemeinbildende Höhere Schulen* [general education high schools, AHS] since 2014/15 (Götz, 2018) and at

Berufsbildende Höhere Schulen [vocational high schools, BHS] since 2015/16 (BMBWF, 2019b). According to the Austrian Ministry of Education, Science and Research, enhancing students' *study ability*, internationally also referred to as *college readiness*, is a purpose of this SRP (BMBWF, 2019b; Singer, 2015). This is consistent with the content of paragraphs 34 and 65 of the School Organization Act, which state that one aim of upper secondary schools is to guide pupils to university readiness (Schulorganisationsgesetz, 1962/2023). That this goal is taken seriously by politicians, shows an announcement made by the then Minister of Education, Heinz Faßmann, that the SRP in mathematics – especially at AHS – should be more oriented around the needs of universities (Austria Presse Agentur, 2020).

Despite the good intention to increase students' study ability, skepticism toward this standardized written examination in mathematics is prevalent among teachers (Pöll, 2017; Singer, 2015), students (Kos, 2020), mathematicians (Kos, 2020; Winkler, 2018), politicians (Thonhauser, 2020), and in various media (Kos, 2020; Pöll, 2017). Götz (2018) responds to some of this criticism in an article, pointing out that most of the criticism expressed occurs without citing empirical evidence. In fact, the research situation regarding the SRP is sparse: besides a few university theses (Behofsits, 2018; Koppenberger, 2022; Kos, 2020; Kutleša, 2018; Plangg & Holzinger, 2018; Pöll, 2017; Reisinger, 2020; Singer, 2015) and theoretical works (Fend, 2020; Fuchs & Kraler, 2020; Götz & Süss-Stepancik, 2018; Haidenthaler, 2019; Sattlberger, 2020; Sattlberger & Steinfeld, 2015; Thonhauser, 2020), hardly any research projects focus empirically on the Austrian SRP in mathematics and its impact. It is therefore barely surprising that, with the exception of LEMMA (Lernstandserhebung, Einstellungen und Motivation zur Mathematik [Assessment of Learning Levels, Attitudes, and Motivation Towards Mathematics]) project (Thaller & Juen-Kretschmer, 2016), no major research programs that intend to investigate the targeted increase in study ability through the SRP in mathematics could be found by the authors.

However, addressing this question would be highly relevant from an educational policy perspective, as "inadequate preparation of pupils for university studies by schools" (van den Berk et al., 2016, p. 17) is often criticized. Increasing study ability is therefore a major policy concern, especially in STEM (science, technology, engineering, and mathematics) subjects (Cramer & Walcher, 2010), where the demand for skilled workers exceeds the supply (Bünning, 2020) and high dropout rates at tertiary education institutions are particularly prevalent (Lumpe, 2019). Furthermore, the neighboring country Germany is struggling with similar difficulties (Bünning, 2020; Lumpe, 2019) and there too is a lively debate about the sense and nonsense of centralized final examinations in mathematics (Leschnig et al., 2022; Neumann et al., 2009). In the meantime, driven mainly by the PISA shock, a so-called *Zentralabitur* [centralized final Abitur examination] in various forms is occurring in 15 out of 16 federal states in Germany. Nevertheless, a nationwide or at least partially overlapping *Zentralabitur* in mathematics is being discussed or even realized in some German federal states (Hoymann, 2011). That this topic is controversially debated is shown by various media contributions, such as an article by the German educationalist Rainer Bölling (Bölling, 2011) in the Frankfurter Allgemeine or a statement by the education economist Lisa Leschnig (Leschnig, 2020).

Since this issue thus seems to be relevant across countries, the further course of this work will investigate if the successful completion of the Austrian SRP in mathematics does indeed increase STEM students' study ability.

BACKGROUND & RESEARCH STATUS

The Austrian SRP

Aiming a comprehensive understanding, this section provides a detailed insight into the structure of the Austrian SRP.

To advance the evolution of the Austrian SRP in mathematics, the AECC-M (Austrian Educational Competence Center Mathematics) was assigned to the conception, preparation, implementation, monitoring, support, and evaluation of a school experiment in 2008 (Dangl et al., 2009). For a detailed description of the design and development process of this standardized examination in mathematics, Dangl et al. (2009) or Sattlberger and Steinfeld (2015) are recommended.

The main result of this process is a standardized written mathematics exam, which is based on a core competence catalog and consists of two parts.

According to the second paragraph of BGBl. II No. 1/2009, competencies are "long-term available cognitive abilities and skills that are developed by learners and that enable them to solve tasks in variable situations successfully and responsibly and to show the related motivational and social readiness". More specifically, core competencies are fundamental and indispensable mathematical skills, which are justified by their high professional and social relevance as well as by curricular demands (Sattlberger & Steinfeld, 2015).

For AHS, the first part of the standardized written mathematics exam contains *type-1* tasks, each of which tests one core competency and is presented with closed-ended answer options like multiple-choice tasks (SRDP, 2021). No skills beyond core competencies are required to solve a type-1 task (SRDP, 2021). The second part contains *type-2* tasks, which are more advanced and complex exercises in which several core competencies must be linked autonomously together (SRDP, 2021). The core competence catalog for AHS includes the content areas *algebra and geometry*, *functional relationships*, *calculus*, as well as *probability and statistics* and can be accessed in detail in SRDP (2022).

The various forms of BHS also have a common core competence catalog in applied mathematics on which the SRP is based on. The common core competence catalog for different BHS covers substantially the same content areas for applied mathematics, with *numbers and measures* additionally mentioned in this case (BMBWF, 2019a). The standardized written exam in applied mathematics also consists of two parts. The first part contains the same type-2-like tasks for all different BHS forms. The second part contains school-specific type-2-like tasks. It should be emphasized that the BHS SRP does not explicitly include type-1 tasks in the written exam for applied mathematics, although a few individual subtasks may occur with narrowly defined answer options.

As previously mentioned, these variants of the SRP in mathematics have been employed as compulsory final exams at AHS since 2014/15 (Götz, 2018) and at BHS since 2015/16 (BMBWF, 2019b). In this context, one aim of this final written examination is to improve students' study ability (BMBWF, 2019b; Singer, 2015). *Study ability* refers to the availability of necessary competencies to successfully pursue and complete an academic degree program. These necessary competencies include, for example, a skillful approach to science (Huber, 1994), well-developed literacy and numeracy skills (Klitzing, 2014), but also robust emotional prerequisites (Schnabl & Kriegler-Kastelic, 2014).

It is assumed that standardized final exams favor pupils study ability, since performance across different federal states and schools can be compared in an objective manner (BMBWF, 2019b). In this way, potential demographic disparities can be uncovered and, at best, evened out. Likewise, deficits in certain content areas can be identified on a broad level and, if necessary, curriculum or teaching adjustments can be made. Additionally, centralized final examinations are a control instrument for the government to ensure that the specified curricular requirements, which according to paragraphs 34 and 65 of the Austrian School Organization Act are supposed to prepare pupils for university studies, are met (Wößmann, 2002). All these mechanisms are intended to help ensure or increase study ability.

Previous Research Results on Study Ability & Centralized Final Examinations

The laudable goal of the Austrian SRP in mathematics raises the question of the actual feasibility of enhancing students' study ability through standardized, centralized final examinations. Empirical evidence on this matter is present within both international and national contexts.

In a research project based on two TIMSS (Trends in International Mathematics and Science Study) data sets from 1995 and 1999 with 447,089 participants, Wößmann (2003) showed that students around the age of 13 from countries with centralized final exams outperform the comparison group in both mathematics and science. A total of 54 countries worldwide participated in the two studies, with 23 countries taking part in both the 1995 TIMSS and the 1999 TIMSS (Wößmann, 2003). Despite 17 control variables for family background, 13 control variables for resource endowment and teacher traits as well as 18 control variables for institutional characteristics of the school system, Wößmann (2003) demonstrated that the achievement gap in favor of students in school systems with centralized exit exams ranged from 35% to 47% of an international standard deviation in test scores, which is substantial. These results reinforce Bishop's (1997) findings, who showed for

the 1995 TIMSS data set that 13-year-old students from countries with centralized exit exams have achievement advantages of the equivalent of 1 U.S. (United States) grade level in mathematics and about 1.2 U.S. grade levels in science. In a more recent study, Leschnig et al. (2022) also demonstrated that centralized final examinations have a positive effect on adults' cognitive abilities, including numeracy literacy.

Klein et al. (2014) point out, however, that the term *centralized final examination* subsumes many inconsistent forms of assessments that differ, for example, in (de)centralized correction. Thus, Klein et al. (2014) conducted a systematic review on the effect of examinations, that must be taken by all students in a given subject at the end of an educational period, which are proctored by an entity external to the school, and comprise the same externally determined tasks for all schools – criteria that are met by the Austrian SRP in mathematics for AHS and partially also for BHS. As a result, there are diverging assumptions and empirical findings on the emotional and motivational experiences of students in relation to centralized final examinations (Klein et al., 2014). One finding that relates specifically to mathematics in Germany is that "students who take central exit exams in mathematics like mathematics less, think it is less easy, and are more likely to find it boring" (Jürges & Schneider, 2010, p. 497). On the other hand, Klein et al. (2014) report predominantly positive results from various investigations concerning achievement gains in mathematics associated with centralized final examinations, although limitations of these results are emphasized.

A study explicitly addressing the relationship between centralized exit exams in upper secondary education and college readiness in the U.S. context was conducted by D'Agostino and Bonner (2009). In the U.S., efforts are underway to strengthen the link between secondary and tertiary educational institutions by defining state standards (D'Agostino & Bonner, 2009). "By 2007, 12 states had aligned their content standards to collegiate expectations, and another 32 states were working towards that goal. Colleges in nine states were using state tests as readiness indicators, and postsecondary schools in 21 additional states were considering the use of exit tests for that purpose" (D'Agostino & Bonner, 2009, p. 26). To investigate whether standardized final exams are indeed predictive of college success, D'Agostino and Bonner (2009) examined the relationship between the *AIMS (Arizona Instrument to Measure Standards) high school mathematics, reading, and writing exam* score and grade point average after the first year at a state university in Arizona for 2,667 students. While a correlation analysis found only a moderate correlation of 0.39 between mathematics scores on the AIMS test and grade point average, further investigation revealed that students who met or surpassed the AIMS mathematics or writing norms achieve a grade point average of C1 (satisfactory)[1] or better with a probability of at least 90% (D'Agostino & Bonner, 2009). **Table 1** (D'Agostino & Bonner, 2009) further shows the proportion of individuals with certain grade point averages after the first university year in relation to their mathematics level on the standardized AIMS test. The achieved AIMS mathematics level consistently indicates a significant ($p<0.01$) positive relationship with grade point average after the first year of university in various regression models (D'Agostino & Bonner, 2009). D'Agostino and Bonner (2009) claim that these study results overcome the limitations of previous studies, which primarily deal with the alignment of centralized exit exams to content standards or entrance test items at tertiary educational institutions. For example, Brown and Conley (2007) found only a moderate fit between the content of state exit tests and necessary mathematical competencies for college readiness, with a few standard areas showing a high alignment. According to D'Agostino and Bonner (2009), such analyses, however, exclude later academic success, measured by grade point average for example.

Table 1. Proportion of individuals with certain average grades after their first year depending on their mathematics level on AIMS standardized test

		Average grade after the first year			
		A	B	C	D or E
Subject … AIMS standards in mathematics	Exceeded	0.64	0.18	0.14	0.04
	Met	0.40	0.27	0.23	0.10
	Approached	0.17	0.26	0.39	0.19
	Fell far below	0.07	0.19	0.49	0.25

In Austria, on the other hand, actual research on the impact of the SRP in mathematics on study skills was planned by the LEMMA project (Thaller & Juen-Kretschmer, 2016). For this project, however, only an interim report was published in 2016, which does not reveal any explicit results concerning study ability and the

[1] Regular grades range from A to E in Arizona, where A means *excellent* and E means *failure*.

Austrian SRP. A diploma thesis by Katrin Kos (2020), which uses data from the LEMMA project, shows that 296 first-year pre-service teachers in mathematics who participated in the SRP have significantly ($p<0.001$) different solution frequencies in a mathematics entrance test than 423 students who did not attend the SRP. While SRP students scored an average of 7.15 out of possible 15 points, non-SRP students scored only 4.37 out of possible 15 points. However, these results should be treated with caution as it is unclear from the data analysis whether and how the comparability of the two groups has been ensured. In addition, it is pointed out that teaching-to-the-test tendencies could be responsible for the better result in the SRP group (Kos, 2020), whereby the mentioned 7.15 points on average do not necessarily speak for a sufficient study ability within this group. We refer to teaching-to-the test as the excessive focus on published or cloned items from the standardized tests during classroom instruction (Volante, 2004). The most harmful form of this practice occurs "when the learning is short term, fading away as soon as the test is over, and when the learning is specific to the items in the test, not generalizable [sic!] to other similar tasks" (Bell, 1994, p. 41). However, it cannot be conclusively clarified how abruptly this forgetting occurs. On one hand, Ebbinghaus' (2011) theory of the forgetting curve assumes that what has been learned – regardless of whether teaching-to-the-test is present or not – is progressively forgotten over time. On the other hand, several studies have demonstrated a positive *testing effect* on memorizing learned content (Carpenter et al., 2008).

Further insights are provided by a dissertation from 2015, which reveals that 234 surveyed AHS mathematics teachers are very skeptical about whether the SRP increases general and subject-relevant study ability (Singer, 2015; K. Singer, personal communication, February 8, 2023). The two used items on study ability are part of a scale comprising 12 items on *quality improvement* through the SRP and are "the new SRP in mathematics contributes to improving general study ability" (Singer, 2015, p. 132) and "the new SRP in mathematics contributes to improving study ability in subject-relevant degree programs" (Singer, 2015, p. 132). These two items have the lowest agreement rate of the whole scale. If 1 corresponds to *disagree* and 4 corresponds to *agree* on a four-point scale, the first item has a mean of 1.78 and the second item has a mean of 1.65, for a total scale mean of 2.175 (Singer, 2015). Of the 234 AHS teachers surveyed, 107 (45.7%) disagreed with the first item on general study ability, 78 (33.3%) somewhat disagreed, 43 (18.4%) somewhat agreed and six (2.6%) agreed (Singer, 2015, personal communication, February 8, 2023). 123 (52.6%) disagreed with the second item on subject-relevant study ability, 57 (32.1%) somewhat disagreed, 30 (12.8%) somewhat agreed and 6 (2.6%) agreed (Singer, 2015, personal communication, February 8, 2023).

Singer (2015) also explicitly addresses the danger of teaching-to-the-test due to the special SRP task formats and could also find evidence for this practice in her research project. Pöll's (2017) research findings also support Singer's (2015) assumption and suggest that teaching-to-the-test is to be expected because of SRP. Moreover, based on 15 interviews with AHS teachers, Pöll (2017) argues in her diploma thesis that teachers would prefer more complex SRP tasks in mathematics for students who are considering university studies.

Furthermore, there are some (not empirically proven) statements by experts who claim that the SRP in mathematics is harmful or at least not beneficial for study ability. Mathematics teacher Tomas Kubelik, for example, writes in a newspaper article for *Die Presse* that he considers study ability as well as general education endangered by the SRP (Kubelik, 2018). Mathematics professor Michael Drmota also reports a steady decline in mathematical knowledge among first-year STEM students and criticizes SRP-focused teaching of mathematics (Taschwer, 2018). Nevertheless it should be noted that Leschnig et al. (2022) were able to find evidence against teaching-to-the-test tendencies through centralized final examinations, although not specifically for the Austrian context.

Overall, however, it can be argued that research is still needed to outline the relationship between the SRP in mathematics and study ability in STEM subjects more precisely. In an article from 2021, Bianca Thaler (2021) summarizes that it remains to be seen what effect the introduction of the SRP will have on study success in Austria.

RESEARCH DESIGN

Research Question

The previous section has clearly shown the need for a more in-depth investigation of the relationship between the introduction of the SRP in mathematics and its intended goal of improving students' study ability. As Bianca Thaler (2021) states it in her outlook question, this research gap is particularly prominent in Austria. Thus, the specific query that will be addressed subsequently is:

Does the SPR in mathematics increase STEM students' study ability in Austria?

Based on the presented national and international research findings as well as the assumed mechanisms of centralized final exams, we expect a positive impact of the Austrian SRP on STEM students' study ability.

Clarification of Relevant Concepts

To answer this research question, it is necessary to clarify what is meant by study ability in general and specifically in relation to STEM subjects.

Huber (1994) describes as study ability the results, i.e., the acquired qualifications, of the preparatory process for dealing with science, which are considered prerequisites for studying in tertiary educational institutions. More specifically, Klitzing (2014) states that study ability consists of "the three basic skills of reading, writing and arithmetic – of course at a high level and embedded in a context of in-depth general education" (Klitzing, 2014, p. 23). Schnabl and Kriegler-Kastelic (2014) emphasize the emotional components of study skills in addition to such cognitive skills. Arnold (2016) mentions, however, that study ability must always be "understood as processual, multidimensional, contextual, and situational" (Arnold, 2016, p. 9). In distinction to *general study ability*, numerous authors therefore address so-called *specific study ability*, for example in relation to a particular discipline or a group of subjects. In this respect, Cramer et al. (2014) describe study ability in mathematics as a neglected topic that implicitly resonates in many educational policy decisions but is not explicitly addressed. The authors specify, for example, that the OECD definition of mathematical literacy given at the beginning of this article does not mention occupation or study preparation in any word (Cramer et al., 2014). Yet, this OECD definition of mathematical literacy is used as the foundation of many education policy reforms – also for developing the Austrian SRP in mathematics (SRDP, 2021). Despite the criticism of the implicit preoccupation with the topic, even Cramer et al. (2014) avoid providing a comprehensive and measurable definition of mathematical study ability that goes beyond initial formulation attempts.

However, for a research project on the specific study ability in physics, Sorge et al. (2016) constructed a multidimensional model of study ability based on previous considerations by several authors. In addition to *academic conditions and general living conditions* as well as *study behavior*, they consider *study ability* to be an essential factor for academic success. Study ability is subdivided into *cognitive*, *subject-related*, *personal,* and *social* dimensions based on Konegen-Grenier's (2002) work. Important components of the cognitive dimension are *analytical skills* and the *ability to abstract*, while the subject-related dimension encompasses the *specialist prior school knowledge* that is necessary to successfully complete a subject-related degree program (Sorge et al., 2016). Whether a study program is considered to be successfully completed can be measured by various indicators such as study satisfaction, competence increase, career success, obtained degree, grades, time to finish the study program, or drop-out (Sorge et al., 2016).

It is precisely this subject-related dimension of study ability that is of interest to this research project since we assume that the SRP in mathematics is focused on this aspect. Hence, the research question can be concretized as follows: Does the SPR in mathematics increase STEM students' subject-related study ability in Austria?

Measurement Instruments & Variables

According to Sorge et al. (2016), the subject-related dimension of study ability is primarily measured by subject school grades or specialized knowledge tests. Thus, one control variable to be considered in our study is the last school grade in mathematics[2].

Furthermore, a mathematical knowledge test for first-year STEM students was created. This test is used in the context of the mathematics bridging course offered to first-year students at the University of Innsbruck. It was developed based on already conducted research on necessary mathematical competencies for STEM studies to ensure content validity regarding subject-related study ability. Specifically, the minimum requirements catalog for mathematics in Baden-Württemberg should be emphasized as a theoretical foundation (cosh, 2012). Empirically, Neumann et al. (2017) conducted a Delphi study on which mathematical competencies are necessary from a college and university perspective to successfully complete STEM degrees. Leaning on these theoretical considerations and empirical results, a competence catalog on which the mathematical test is based was created. In the context of thirteen bachelor's theses, competence-oriented tasks with narrowly defined answer options were developed and reviewed by a mathematician and mathematics education researcher. To obtain feedback, the designed exercises were then submitted to an expert from the Department III/6f of the Federal Ministry of Education, Science and Research, which is responsible for creating the SRP tasks for AHS. Finally, the test was piloted and revised at the University of Innsbruck during the winter term 2019/20. The final test comprises 129 type-1-like items, 20 of which are identical or very similar to actual SRP items. The test yields a WLE (weighted-likelihood-estimates) reliability of 0.921 and an EAP (expected a posteriori) reliability of 0.925, which is considered very good for performance tests (Gäde et al., 2020). A multidimensional Rasch model for the four content areas *algebra and geometry*, *functional relationships*, *calculus,* and *probability and statistics* shows a significantly ($p<0.001$) better model fit in a likelihood ratio test but lower WLE and EAP reliabilities[3] in each dimension compared to the unidimensional model. Moreover, the narrowly defined answer options of the developed tasks ensure objectivity as the correction process can be automated.

Going beyond content validity and reliability, in accordance with D'Agostino and Bonner (2009) criticism of purely content-oriented alignment, we also ensured that the developed test is predictive of study success measured by ECTS (European Credit Transfer and Accumulation System) credits earned per semester and dropout. In fact, a highly significant ($p<0.001$), positive correlation between the solution frequency in the test and ECTS credits achieved per semester (Tscholl, 2023; Tscholl et al., Manuscript submitted for publication) as well as the probability of remaining in the chosen STEM degree program could be detected (Tscholl, 2023). Therefore, the primary output variable for this research project is the subject-related dimension of study ability measured by the solving frequency of 109 non-SRP tasks in the developed assessment to avoid teaching-to-the-test effects in the SRP group as good as possible. This variable is referred to as *subject-related study ability* in the following.

The primary input variable is self-reported participation in the SRP in mathematics, which can be expressed in a binary manner. Control variables are, in addition to the already mentioned last school grade in mathematics, gender, STEM field, age, difficulties while solving the mathematical test, attended school type, mathematical self-concept, mathematical self-efficacy, and the possible previous attendance in a STEM degree program. All control variables were obtained by means of questionnaires. The last school grade in mathematics was surveyed categorically in the Austrian grading system from 1 (=very good) to 5 (= not sufficient). Regarding gender, the categories *male*, *female*, and *diverse* were offered for selection. However, the latter category was removed from the sample due to its very small sample size. The age was obtained in years and difficulties solving the mathematical test include the answer options *lack of motivation/concentration*, *lack of practice*, and *no technology use*. The attended school type covers the categories *AHS* and *BHS,* while the question about previously attended STEM studies could be answered with *yes* or *no*.

[2] 1 means *very good*, 2 means *good*, 3 means *satisfactory*, 4 means *sufficient*, & 5 mean *non-sufficient*.

[3] **Algebra & Geometry:** WLE reliability 0.858; EAP reliability 0.897, **Functional Relationships:** WLE reliability 0.619; EAP reliability 0.819, **Calculus:** WLE reliability 0.64; EAP reliability 0.829, & **Probability & Statistics:** WLE reliability 0.623; EAP reliability 0.816.

Mathematical self-concept and mathematical self-efficacy were assessed using a five-point Likert scale developed by the Federal Ministry of Education, Science and Research.

Sample

The data for the described variables were collected from 217 Austrian first-year[4] STEM students who participated in the mathematics bridging course at the University of Innsbruck between 2020 and 2023 (three cohorts). The participants solved the mathematical test at the beginning of the semester before any instruction and mostly without technological or other aids. Further data were voluntarily provided by the students with permission to use their data for research purposes.

Of the 217 students, only 17 did not participate in the SRP in mathematics. This circumstance leads to a serious imbalance with respect to the experimental group (SRP) and the control group (non-SRP). To account for this imbalance and to create comparable conditions between groups, a 1:1 nearest neighbor matching without replacement based on probit regression is performed in *R* with the package *MatchIt* on all available control variables. This matching algorithm could assign 17 people from the experimental group to the available 17 people from the control group in a way that they do not significantly differ from each other in the control variables (see **Table A1** in **Appendix A**).

However, the sample size is reduced drastically through this procedure. To increase the sample slightly and to ensure its robustness, the matching algorithm was repeated several times, prioritizing a different control variable in each run. This process created five additional experimental groups, which show no significant differences in the remaining control variables. More specifically, as shown in **Table 2**, six sample cases emerge in terms of sample composition.

Table 2. Overview of six different sample cases

SC	Prioritized control variable	SS	Remarks
1	/	17	No significant differences in control variables (**Table A1** in **Appendix A**)
2	Age	11	No significant differences in control variables (**Table A2** in **Appendix A**)
3	Gender	17	No significant differences in control variables (**Table A3** in **Appendix A**)
4	First STEM semester	17	No significant differences in control variables (**Table A4** in **Appendix A**)
5	Lack of practice while solving the test	17	No significant differences in control variables (**Table A5** in **Appendix A**)
6	Attended school type	17	No significant differences in control variables (**Table A6** in **Appendix A**)

Note. SC: Sample case & SS: Sample size per group

The algorithm actually used 22 of the 200 people from the experimental group for matching. A more detailed insight into the sample composition can be found in **Table A1** to **Table A6** in **Appendix A**. **Table A7** in **Appendix A** provides a detailed overview of which people are in which sample case. Nine people of the experimental group are assigned to each of the sample cases, which shows that the matching is robust even when prioritizing different variables.

Statistical Methods

Since we matched people from the experimental group to the control group, a quasi-experimental study design is present.

After checking the requirements, a one-tailed two-sample Wilcoxon rank sum test is applied group-wise for each sample case to the subject-related study ability variable. For each sample case, the null hypothesis is that the experimental group, i.e., SRP group, and the control group, i.e., the non-SRP group, do not differ in their median subject-related study ability. Accordingly, the alternative hypothesis is that the experimental group has a higher median subject-related study ability than the control group.

If no significant results are found, a post hoc power analysis will subsequently be performed. To calculate the minimum sample size at a given power, the R package *MKpower* is used, which performs power analyses based on Monte-Carlo simulations. We want to emphasize that this approach does not have the purpose of justifying non-significant results, but mainly to act as a guideline for future research projects.

Finally, comparative correlation analyses are used as robustness tests to detect whether students of the experimental groups have been explicitly trained to solve specific SRP tasks. If a competence is present

[4] In this context, *first-year* means that the students are in the first semester of the STEM program with which they enrolled in the mathematics bridging course with which they enrolled for mathematics bridging course.

independently of such teaching-to-the-test, the correlation between the solving frequency of the 20 SRP tasks and the remaining 109 assessment tasks, i.e., subject-related study ability, should be equally strong in the experimental and control groups. If, on the other hand, teaching-to-the-test regarding specific SRP tasks occurred, there should be a less strong correlation between the solving frequency of the 20 SRP tasks and the remaining 109 assessment tasks in the experimental groups. For this purpose, one-tailed Fisher's Z within the *cocor* package in *R* is used to test whether the Pearson correlation between the solving frequency of the 20 SRP tasks and the 109 assessment tasks is smaller in the experimental group than in the control group. Furthermore, visual analysis will be used to check whether the difference in the solving frequency between the 20 SRP tasks and the 109 assessment tasks shows deviating patterns between the experimental and control groups.

RESULTS

Group comparisons regarding subject-related study ability via one-tailed two-sample Wilcoxon rank sum test do not yield significant results at a 5% significance level (see **Table 3**). The null hypothesis, being that the experimental group and control group do not differ in the median solving frequencies, must be retained accordingly. In addition, all sample cases except sample case 2 show low effect sizes (see **Table 3**). Based on the mean values and standard deviations of the solving frequencies per group (see **Table 3**), a post-hoc power analysis is conducted.

Table 3. Results of one-tailed Wilcoxon rank sum test regarding subject-related study ability

	SC	Control group	Experimental group	Estimated location shift with lower limit for 95% CI	Cohen's d	p-value
Subject-related study ability	1	Median with 25% & 75% quantiles: 0.4 (0.28 to 0.57) Mean with SD: 0.44 (0.18)	Median with 25% & 75% quantiles: 0. 45 (0.37 to 0.63) Mean with SD: 0.49 (0.16)	0.062 (-0.033)	0.278	0.143
	2	Median with 25% & 75% quantiles: 0.4 (0.29 to 0.43) Mean with SD: 0.39 (0.12)	Median with 25% & 75% quantiles: 0.45 (0.37 to 0.6) Mean with SD: 0.49 (0.15)	0.079 (-0.009)	0.667	0.074
	3	Median with 25% & 75% quantiles: 0.4 (0.28 to 0.57) Mean with SD: 0.44 (0.18)	Median with 25% & 75% quantiles: 0.4 (0.37 to 0.63) Mean with SD: 0.48 (0.16)	0.046 (-0.042)	0.222	0.185
	4	Median with 25% & 75% quantiles: 0.4 (0.28 to 0.57) Mean with SD: 0.44 (0.18)	Median with 25% & 75% quantiles: 0.45 (0.37 to 0.63) Mean with SD: 0.5 (0.16)	0.065 (-0.028)	0.333	0.114
	5	Median with 25% & 75% quantiles: 0.4 (0.28 to 0.57) Mean with SD: 0.44 (0.18)	Median with 25% & 75% quantiles: 0.39 (0.28 to 0.63) Mean with SD: 0.5 (0.15)	0.064 (-0.022)	0.333	0.107
	6	Median with 25% & 75% quantiles: 0.4 (0.28 to 0.57) Mean with SD: 0.44 (0.18)	Median with 25% & 75% quantiles: 0.45 (0.37 to 0.63) Mean with SD: 0.49 (0.16)	0.064 (-0.033)	0.278	0.124

Note. SC: Sample case; SD: Standard deviation; & CI: Confidence interval

Table 4 displays the results of this power analysis for a one-tailed two-sample Wilcoxon rank sum test obtained by the *R* package *MKpower*. As **Table 4** shows, with the deviations estimated in **Table 3**, between 30 and 260 people per group would be needed to detect a significant effect with a statistical power of at least 80%. Since a (non-significant) tendency in favor of the experimental groups is apparent in **Table 3**, correlation analyses are performed as robustness tests regarding teaching-to-the-test practices.

Table 4. Minimal sample size per group for each sample case to retain a statistical power of at least 80% with one-tailed two-sample Wilcoxon rank sum test

Sample case	Minimum sample size per group for a statistical power of at least 80%
1	170
2	30
3	260
4	120
5	130
6	140

As **Table 5** shows, the correlation between the solving frequency of the 109 assessment tasks and the 20 SRP tasks is significantly lower in the experimental group for four of the six tested sample cases.

Table 5. Results of one-tailed correlation analysis for each sample case

Sample case	Correlation control group	Correlation experimental group	p-value
1	0.94	0.78	0.035
2	0.88	0.82	0.341
3	0.94	0.83	0.082
4	0.94	0.79	0.046
5	0.94	0.79	0.041
6	0.94	0.78	0.038

The patterns of this unexpected deviation are examined further in the bottom row of **Figure 1** and **Figure 2**:

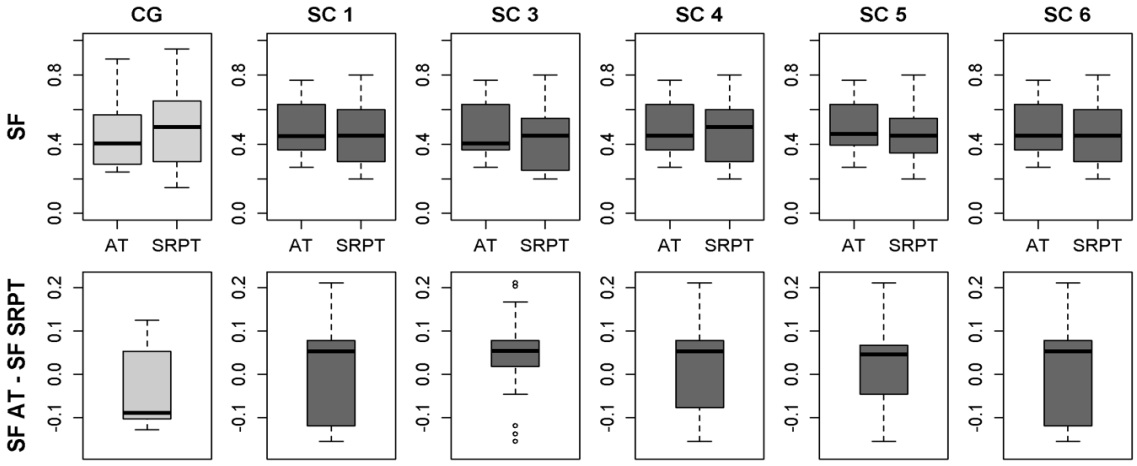

Figure 1. First row shows solving frequency (SF) of 109 assessment tasks (AT) & 20 SRP tasks (SRPT) for control group (CG) with 17 students (light gray) & experimental groups for sample cases (SC) 1, 3, 4, 5, & 6 (dark gray); second row shows distributions of person-wise differences in solving frequency of assessment tasks & SRP tasks for control group & experimental groups for sample cases 1, 3, 4, 5, & 6 (Source: Authors' own elaboration)

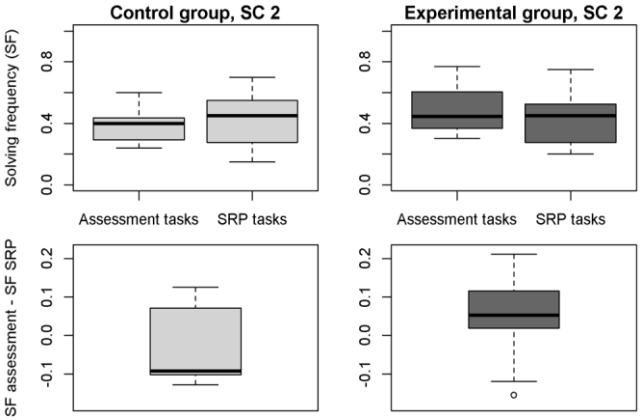

Figure 2. First row shows solving frequency (SF) of 109 assessment tasks & 20 SRP tasks for control group with 11 students (light gray) & experimental group for sample case (SC) 2 (dark gray); second row shows distributions of person-wise differences in solving frequency of assessment tasks & SRP tasks for control group & experimental group for sample case 2 (Source: Authors' own elaboration)

There is no indication that the experimental groups are particularly good in solving the 20 SRP tasks compared with the 109 assessment tasks. Rather the opposite, the control groups show negative median values in the bottom row boxplots of **Figure 1** and **Figure 2** (which means that the person-wise solving frequency of SRP tasks tends to be above that of the assessment tasks), while the median values for all experimental groups lie above 0 (which means that the person-wise solving frequency of the assessment tasks tends to be above that of the SRP tasks).

DISCUSSION & LIMITATIONS

The results of the statistical analysis in the previous section suggest that participation in the SRP in mathematics does not lead to a significantly higher subject-related study ability in STEM subjects. This result contradicts the findings of previous studies and systematic reviews, which show a positive relationship between standardized final examinations and performance in mathematics and science (Bishop, 1997; Klein et al., 2014; Wößmann, 2003). Specifically, these results also oppose Kos' (2020) findings that first-year mathematics teacher trainees who participated in the SRP in mathematics perform better in a mathematics test than freshmen who did not participate in the SRP in mathematics. Nonetheless, it can be argued that a positive performance trend in favor of the experimental groups is visible in the different sample cases displayed in **Table 3** – even if the effect is mostly small. This trend is consistent with the presented research results regarding study ability and centralized final exams, although previous studies have found significant and larger effects (Bishop, 1997; Wößmann, 2003). Furthermore, a positive relationship is to be expected due to the presumed mechanisms: It is assumed that centralized final examinations allow objective comparisons and are therefore a control instrument for the government to adapt and enforce curricular requirements (BMBWF, 2019b; Wößmann, 2002).

It is particularly noteworthy that this found trend is most pronounced (and even nearly significant) for sample case 2, which favored the students' age in the matching process. Thus, it cannot be argued with Ebbinghaus' (2011) theory on the forgetting curve that the control group, which tends to be older in the other sample cases, has already forgotten more school knowledge and therefore performs tendentially worse regarding subject-related study ability. In addition, no teaching-to-the-test tendencies could be detected in the experimental groups, although concerns in this regard are frequently mentioned (Kos, 2020; Pöll, 2017; Singer, 2015; SRDP, 2022; Taschwer, 2018). As the bottom rows in **Figure 1** and **Figure 2** show, unlike the control groups, the experimental groups tend to solve SRP tasks worse than the assessment tasks. Thus, if teaching-to-the-test took place, the forgetting process must have occurred very abruptly and extensively right after the SRP, as described by Bell (1994). However, the tendentially better overall performance of the experimental groups in the subject-related study ability test stands against this hypothesis.

Unfortunately, the sample size of the sample cases is too small to provide reliable answers to the research question, as confirmed by the post hoc power analyses (see **Table 4**). But even with non-significant results, the study's findings cannot support Kubelik's (2018) assertion that study ability is at risk due to the SRP in mathematics.

Nevertheless, our findings should be treated with caution. The already mentioned major weakness of the analysis is certainly the small sample size. To counteract this flaw, different experimental groups were matched to the existing control group. However, since the initial matching turned out to be robust when prioritizing different control variables, the matched samples overlap to some extent. This also undermines the conclusion that the applied tests for the various sample cases exhibit a consistent trend. Additionally, the samples within each sample case consist primarily of BHS graduates whose written SRP in mathematics contains no or hardly any type-1 tasks. Therefore, they may experience less teaching-to-the-test regarding this special task type compared to AHS graduates. For a sample consisting exclusively of AHS graduates, the robustness test results (see **Table 5**, **Figure 1**, and **Figure 2**) could therefore be different. Furthermore, the SRP tasks used in the teaching-to-the-test analysis deviate slightly from the original tasks, for example by using different numbers. Therefore, the correlation analysis to measure teaching-to-the-test tendencies in relation to specific SRP tasks may be biased.

Another limitation concerning the sample is a possible self-selection bias. Since the initial 217 individuals voluntarily enrolled in the mathematics bridging course, the selection may not be representative of the entire student population. However, it can be seen from **Table A1** to **Table A6** (**Appendix A**) that the subjects in the remaining sample cases show at least a broader range of mathematics grades. Therefore, it cannot be assumed that the sample cases consist exclusively of high-performing individuals, as may be expected through voluntary participation in the bridging course (Büchele, 2020; Gerdes et al., 2022; Tieben, 2019).

The primary input variable *SRP participation in mathematics* and the primary output variable *subject-related study ability* also have weaknesses. On one hand, looking exclusively at SRP participation may be insufficient

to adequately account for other changes in the school system unleashed by the introduction of the SRP. Thus, participation in SRP may only be a proxy for other influential developments, e.g., regular use of technology in mathematics lessons. On the other hand, subject-related study ability measured by the solving frequency in the 109 assessment tasks is also subject to inaccuracies. For example, some students may have solved the assessment improperly with technological or other aids. However, it should be noted that only 53 of the 217 first-year students, respectively a maximum of two people in the experimental and control groups of each sample case, retrospectively stated that the lack of technology use was the biggest obstacle in solving the assessment tasks (see **Table A1** to **Table A6** in **Appendix A** for a more detailed insight).

Finally, it should be emphasized that the average subject-related study ability score measured by the assessment solving frequency lies between 0.39 and 0.50 (see **Table 3**), which is humbling. However, the assessment includes 109 tasks, some of which are very demanding and go beyond mathematical school knowledge. Yet, the average solving frequencies of the 20 SRP tasks that do not go beyond school curriculum content also lie between 0.42 and 0.48 for the experimental and control groups. If this is used as a general estimate, it can be assumed that only about every second SRP task in mathematics can be solved by first-year STEM students with a high school diploma.

SUMMARY & OUTLOOK

Since 2014/15 for AHS and since 2015/16 for BHS, a standardized written final examination in mathematics or applied mathematics has been compulsory for nearly all pupils at the upper secondary level in Austria (BMBWF, 2019b; Götz, 2018). While the Federal Ministry of Education, Science and Research states that this SRP is intended to increase pupils' (subject-related) study ability (BMBWF, 2019b; Singer, 2015), the empirical research base in this regard is scarce in Austria. Although international studies indeed suggest that standardized school-leaving examinations have a positive impact on mathematics or science performance (Bishop, 1997; Klein et al., 2014; Wößmann, 2003), there is skepticism about whether the SRP actually increases subject-related study ability in Austria (Kubelik, 2018; Pöll, 2017; Singer, 2015; Taschwer, 2018). In this respect, Kos' (2020) findings show that first-year pre-service teachers who have participated in the SRP in mathematics outperform their peers in a mathematics test. However, the reason for this result could be teaching-to-the-test (Kos, 2020; Singer, 2015).

To contribute to the thin research base, the current study draws on 217 first-year STEM students who took part in the mathematics bridging course at the University of Innsbruck between 2020/21 and 2022/23.

Since only 17 of the 217 students did not participate in the SRP in mathematics, the sample was reduced using 1:1 nearest neighbor matching without replacement based on probit regression. This matching process was performed six times, with no control variable being prioritized for matching in the first run, the control variable *age* in the second run, *gender* in the third run, *first STEM semester* in the fourth run, *lack of practice while solving the test* in the fifth run, and the *attended school type* in the sixth run. This matching process results in six different sample cases, each containing 17 individuals in the control and in the experimental group (sample cases 1, 3, 4, 5, and 6) and 11 individuals in the control and in the experimental group (sample case 2), respectively.

On this reduced sample cases, no significant differences between the experimental and control groups' subject-related study ability were found via a one-tailed two-sample Wilcoxon rank sum test – although a tendency in favor of the experimental groups emerged. Subject-related study ability was measured by a mathematics test, which was designed based on prior theoretical and empirical results (cosh, 2012; Neumann et al., 2017). Additionally, performance in the developed test is predictive of academic success measured by dropout and ECTS credits achieved per semester (Tscholl, 2023; Tscholl et al., Manuscript submitted for publication).

Furthermore, post hoc power analyses on all sample cases show that the sample size would have to be (much) larger to prove significant effects with a power of at least 80% (see **Table 4**). In addition, robustness tests detect no teaching-to-the-test effects in the experimental groups, which could bias the results (see **Table 5**, **Figure 1**, and **Figure 2**).

In conclusion, neither an increase in subject-related study ability within the experimental groups could be demonstrated nor are teaching-to-the-test tendencies visible. This result, which contradicts previous studies (Bishop, 1997; D'Agostino & Bonner, 2009; Klein et al., 2014; Kos, 2020; Singer, 2015; Wößmann, 2003), could arise due to the small sample size and the special sample, which mainly consists of BHS graduates.

Based on the insights provided by **Table 4**, future studies should aim for a sample size of about 150 students per research group and consider the school type-specific differences between AHS and BHS concerning the SRP. Ideally, an experimental study design is aimed to exclude any selection biases or other distorting influences of control variables in the best possible way. To mitigate the danger that found effects of the SRP on study ability are merely proxy artifacts, future studies could target a study design with AHS students of the 2014/15 graduation class or with BHS students of the 2015/16 graduation class. Within this framework, a retrospective assessment of study ability could be realized by evaluating the successful completion of STEM studies. Notably, these cohorts represent the pioneering groups, where the SRP in mathematics was made compulsory. Consequently, educators had limited opportunities to effectively align their pedagogical strategies to the SRP, such as implementing modified technology usage within the classroom. Furthermore, the challenges of teaching-to-the-test influences are minimized, given that instructors themselves had only a rudimentary grasp of what to expect from the SRP in mathematics at the end of the school year. In any case, teaching-to-the-test tendencies should be considered in future research projects, as these are often suspected (Kos, 2020; Pöll, 2017; Singer, 2015; Taschwer, 2018), even if our study does not find any evidence for this – at least in a sample consisting mainly of BHS graduates.

Author contributions: PT: wrote manuscript & performed data analysis; **FS & TH:** responsible for development of test instrument used to measure study ability & supervised data analysis & writing process. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: The authors declare that the used data were collected in accordance with Article 13 of the Austrian General Data Protection Regulation and with the written consent of the adult participants.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Arnold, E. (2016). Vorwort [Preface]. In I. van den Berk, K. Petersen, K. Schultes, & K. Stolz (Eds.), *Universitätskolleg-Schriften: Studierfähigkeit: Theoretische Erkenntnisse, empirische Befunde und praktische Perspektiven* [*University college writings: Study ability: Theoretical insights, empirical findings and practical perspectives*] (pp. 9-13).

Austria Presse Agentur. (2020). Mathe-Zentralmatura an AHS wird auf neue Beine gestellt [Mathematics Central Matura at AHS is being put on a new footing]. *Tiroler Tageszeitung*. https://www.kleinezeitung.at/politik/innenpolitik/5835905/Personalwechsel_MatheKlausur-an-AHS-wird-auf-neue-Beine-gestellt

Behofsits, A. (2018). *Mathematik und Sprache: Vorbereitung auf die kompetenzorientierte standardisierte Reifeprüfung durch sprachsensiblen Mathematikunterricht* [*Mathematics and language: Preparation for the competency-oriented standardized matriculation examination through language-sensitive mathematics lessons*] [Master's thesis, Karl-Franzens-Universität Graz].

Bell, A. (1994). Teaching for the test. In M. Selinger (Ed.), *Teaching mathematics* (pp. 41-46). Routledge.

Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *The American Economic Review, 87*(2), 260-264.

BMBWF. (2019a). Mathematische Grundkompetenzen im gemeinsamen Kern: Gültig ab den Matura-Prüfungsterminen 2017/2018 [Basic mathematical skills in the common core: Valid from the Matura examination dates 2017/2018]. *Bundesministerium für Bildung, Wissenschaft und Forschung* [*Federal Ministry of Education, Science and Research*]. https://www.matura.gv.at/fileadmin/user_upload/downloads/Begleitmaterial/AM/srdp_am_kompetenzen_2018_teil_a-2019-09-05.pdf

BMBWF. (2019b). Die standardisierte Reife- und -Diplomprüfung: …für höchste Qualität an Österreichs Schulen [The standardized maturity and diploma examination: …for the highest quality in Austria's schools]. *Bundesministerium für Bildung, Wissenschaft und Forschung* [*Federal Ministry of Education, Science and Research*]. https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura/srdp.html

Bölling, R. (2011). Auch ein Zentralabitur bürgt nicht für Qualität [Even a central high school diploma does not guarantee quality]. *Frankfurter Allgemeine*. https://www.faz.net/aktuell/karriere-hochschule/campus/hochschulreife-auch-ein-zentralabitur-buergt-nicht-fuer-qualitaet-1590722.html

Brown, R. S., & Conley, D. T. (2007). Comparing state high school assessments to standards for success in entry-level university courses. *Educational Assessment, 12*(2), 137-160. https://doi.org/10.1080/10627190701232811

Büchele, S. (2020). Bridging the gap–How effective are remedial math courses in Germany? *Studies in Educational Evaluation, 64*, 100832. https://doi.org/10.1016/j.stueduc.2019.100832

Bünning, F. (2020). Fachkräftemangel und Handlungsfelder für die technische Bildung [Shortage of skilled workers and areas of action for technical education]. In F. Bünning, M. Dick, R. W. Jahn, & A. Seltrecht (Eds.), *Berufsbildung, Arbeit und Innovation: Zwischen Ingenieurpädagogik, Lehrkräftebildung und betrieblicher Praxis eine Festschrift für Klaus Jenewein* [*Vocational training, work and innovation: A commemorative publication for Klaus Jenewein between engineering education, teacher training and operational practice*] (pp. 117-126). W. Bertelsmann Verlag. https://doi.org/10.3278/6004727w

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. https://doi.org/10.3758/MC.36.2.438

cosh. (2012). *Mindestanforderungskatalog Mathematik: Der Hochschulen Baden-Württembergs für ein Studium von MINT- oder Wirtschaftsfächern. Ergebnis einer Fachtagung am 5. Juli 2012 in Esslingen* [*Minimum requirements catalog for mathematics: Baden-Württemberg universities for studying MINT or business subjects. Result of a specialist conference on July 5, 2012 in Esslingen*]. https://lehrerfortbildung-bw.de/u_matnatech/mathematik/bs/bk/cosh/katalog/makv20b_ohne_leerseiten.pdf

Cramer, E., & Walcher, S. (2010). Schulmathematik und Studierfähigkeit [School mathematics and study skills]. *Mitteilungen Der Deutschen Mathematiker-Vereinigung* [*Announcements from the German Association of Mathematicians*]*, 18*, 110-114.

Cramer, E., Walcher, S., & Wittich, O. (2014). Studierfähigkeit im Fach Mathematik: Anmerkungen zu einem vernachlässigten Thema [Ability to study mathematics: Notes on a neglected topic]. In S. Lin-Klitzing, D. Di Fuccia, & R. Stengl-Jörns (Eds.), *Gymnasium-Bildung-Gesellschaft: Abitur und Studierfähigkeit: Ein interdisziplinärer Dialog* [*High school education society: Abitur and ability to study: An interdisciplinary dialogue*] (pp. 163-182). Klinkhardt.

D'Agostino, J. V., & Bonner, S. M. (2009). High school exit exam scores and university performance. *Educational Assessment, 14*(1), 25-37. https://doi.org/10.1080/10627190902816223

Dangl, M., Fischer, R., Heugl, H., & Kröpfl, B. (2009). *Das Projekt "Standardisierte schriftliche Reifeprüfung aus Mathematik": Sicherung von mathematischen Grundkompetenzen* [*The "standardized written matriculation examination in mathematics" project: Securing basic mathematical skills*]. https://www.aau.at/wp-content/uploads/2017/10/sRP-M_September_2009-2.pdf

Ebbinghaus, H. (2011). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [*On memory: Studies in experimental psychology*]. Verlag von Duncker & Humber.

Fend, H. (2020). Das österreichische Bildungswesen vor einer Zeitwende? Kontrollschub oder Qualitätsschub: Wohin führt die neue "Testungsmaschine"? [The Austrian education system facing a turning point? Control boost or quality boost: Where will the new "testing machine" lead?] In U. Greiner, C. Wiesner, C. Schreiner, & F. Hofmann (Eds.), *Bildungsstandards: Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* [*Educational standards: Competence orientation, task culture and quality development in the school system*] (pp. 561-576). Waxmann Verlag.

Fuchs, K., & Kraler, C. (2020). Bildungsstandards und Aufgabenkultur im Mathematikunterricht [Educational standards and task culture in mathematics lessons]. In U. Greiner, C. Wiesner, C. Schreiner, & F. Hofmann (Eds.), *Bildungsstandards: Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* [*Educational standards: Competence orientation, task culture and quality development in the school system*] (pp. 482-500). Waxmann Verlag.

Gäde, J. C., Schermelleh-Engel, K., & Werner, C. S. (2020). Klassische Methoden der Reliabilitätsschätzung [Classic methods of reliability estimation]. In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* [*Test theory and questionnaire construction*] (p. 305-334). Springer. https://doi.org/10.1007/978-3-662-61532-4_14

Gerdes, A., Halverscheid, S., & Schneider, S. (2022). Teilnahme an mathematischen Vorkursen und langfristiger Studienerfolg. Eine empirische Untersuchung [Participation in preliminary mathematical courses and long-term academic success. An empirical study]. *Journal Für Mathematik-Didaktik* [*Journal for Mathematics Didactics*]*, 43*(2), 377-403. https://doi.org/10.1007/s13138-021-00194-3

Götz, S. (2018). Eine echte Teilmenge: Zum Konzept der standardisierten schriftlichen Reifeprüfung in Mathematik an AHS [A real subset: On the concept of the standardized written matriculation examination in mathematics at AHS]. *Mathematik Im Unterricht* [*Mathematics in the Classroom*]*, 9*, 15-28.

Götz, S., & Süss-Stepancik, E. (2018). Vom Testen zum (kompetenten) Lernen im Mathematikunterricht [From testing to (competent) learning in mathematics lessons]. *R&E-Source*.

Haidenthaler, W. (2019). *SRDP Mathematik spricht Deutsch. Themenprogramm: Kompetenzen im mathematischen und naturwissenschaftlichen Unterricht* [*SRDP mathematics speaks German. Thematic program: Competencies in mathematics and science lessons*]. https://www.imst.ac.at/files/projekte/2019/berichte/2019_Langfassung_Haidenthaler.pdf

Hoymann, T. (2011). *Umdenken nach dem PISA-schock* [*Rethinking after the PISA shock*]. Tectum-Der Wissenschaft.

Huber, L. (1994). Nur allgemeine Studierfähigkeit oder doch allgemeine Bildung? Zur Wiederaufnahme der Diskussion über "Hochschulreife" und die Ziele der Oberstufe [Just general study skills or general education? To resume the discussion about "university readiness" and the goals of high school]. *Die Deutsche Schule* [*The German School*]*, 86*(1), 12-26.

Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance … but take the fun out of mathematics. *Journal of Population Economics, 23*(2), 497-517. https://doi.org/10.1007/S00148-008-0234-3

Klein, E. D., Krüger, M., Kühn, S. M., & van Ackeren, I. (2014). Wirkungen zentraler Abschlussprüfungen im Mehrebenensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde und Forschungsdesiderata [Effects of central final examinations in the multi-level school system. An interim assessment of international and national findings and research desiderata]. *Zeitschrift Für Erziehungswissenschaft* [*Journal of Educational Science*]*, 17*(1), 7-33. https://doi.org/10.1007/s11618-014-0479-4

Klitzing, H. G. (2014). Studierfähigkeit–die schulische und verbandspolitische Sicht [Ability to study–the school and association policy perspective]. In S. Lin-Klitzing, D. Di Fuccia, & R. Stengl-Jörns (Eds.), *Gymnasium–Bildung–Gesellschaft: Abitur und Studierfähigkeit: Ein interdisziplinärer Dialog* [*High school–education–society: Abitur and ability to study: An interdisciplinary dialogue*] (pp. 23-26). Klinkhardt.

Konegen-Grenier, C. (2002). *Studierfähigkeit und Hochschulzugang* [*Ability to study and university entrance*]. Dt. Inst.-Verl.

Koppenberger, C. (2022). *Lernvideos als Vorbereitung auf die schriftliche Reifeprüfung im Fach Mathematik: Fünf Lernvideos zum Thema Änderungsmaße im Vergleich* [*Learning videos as preparation for the written matriculation examination in mathematics: Five learning videos on the subject of change measures in comparison*] [Bachelor thesis, Private Pädagogische Hochschule der Diözese Linz].

Kos, K. (2020). *Einfluss der standardisierten Reifeprüfung auf die Leistungen von Studienanfängerinnen und Studienanfängern im Lehramt Mathematik* [*Influence of the standardized school leaving examination on the performance of new students in mathematics teaching*] [Diploma thesis, Karl-Franzens-Universität Graz].

Kubelik, T. (2018). Neue Mathematik-Matura: Kolossal banal [New mathematics matura: Colossally banal]. *Die Presse*. https://www.diepresse.com/5405644/neue-mathematik-matura-kolossal-banal?utm_campaign=Echobox&utm_medium=Social&utm_source=Facebook&xtor=CS1-15&fbclid=IwAR1ZVUjeOMTT20gtD5_x2pyM4ZElwoqcFik28U-oOVV7SWrt9RYp8gdt5Gk&pw-overlay-registration=started

Kutleša, P. (2018). *Standardisierte Reifeprüfung: Ein Vergleich ausgewählter Staaten mit dem Schwerpunkt auf der österreichischen standardisierten Reifeprüfung im Fach Mathematik* [*Standardized school leaving examination: A comparison of selected countries with the focus on the Austrian standardized school leaving examination in mathematics*] [Master's thesis, Karl-Franzens-Universität Graz].

Leschnig, L. (2020). Das bundesweite Zentralabitur–ein richtiger Ansatz, aber kein Allheilmittel [The nationwide central high school diploma–A right approach, but not a panacea]. *IAB*. https://www.iab-forum.de/das-bundesweite-zentralabitur-ein-richtiger-ansatz-aber-kein-allheilmittel/

Leschnig, L., Schwerdt, G., & Zigova, K. (2022). Central exams and adult skills: Evidence from PIAAC. *Economics of Education Review, 90*, 102289. https://doi.org/10.1016/j.econedurev.2022.102289

Lumpe, M. (2019). Studienabbruch in den MINT-Fächern: Fallstudien an der Universität Potsdam und mögliche Folgerungen [Dropping out of studies in the MINT subjects: Case studies at the University of Potsdam and possible consequences]. In W. Schubarth, S. Mauermeister, F. Schulze-Reichelt, & A. Seidel (Eds.), *Potsdamer Beiträge zur Hochschulforschung: Alles auf Anfang! Befunde und Perspektiven zum Studieneingang* [*Potsdam contributions to university research: Everything at the beginning! Findings and perspectives on the start of the study*] (pp. 177-194). Universitätsverlag Potsdam.

Neumann, I., Pigge, C., & Heinze, A. (2017). *Welche mathematischen Lernvoraussetzungen erwarten Hochschullehrende für ein MINT-Studium?* [*What mathematical learning requirements do university teachers expect for a MINT degree?*] https://www.telekom-stiftung.de/sites/default/files/files/media/publications/MaLeMINT_Broschu%CC%88re_Korr.%20Version%20Mai%202018.pdf

Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen [Comparability of high school diploma achievements]. *Zeitschrift Für Erziehungswissenschaft* [*Journal of Educational Science*]*, 12*(4), 691-714. https://doi.org/10.1007/s11618-009-0099-6

OECD. (2003). The PISA 2003 assessment framework–Mathematics, reading, science and problem solving knowledge and skills. *Organization for Economic Co-Operation and Development*. https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33694881.pdf

Plangg, S., & Holzinger, J. (2018). *Schriften zur Didaktik der Mathematik und Informatik an der Universität Salzburg: Aktuelle Themen fachdidaktischer Forschung: Mathematikunterricht im Wandel-Eine fachdidaktische Analyse-Kommunikation im Mathematikunterricht* [*Writings on the didactics of mathematics and computer science at the University of Salzburg: Current topics in subject-didactic research: Mathematics teaching in transition–A subject-didactic analysis-communication in mathematics lessons*]. Shaker.

Pöll, J. (2017). *Ängste, Schwierigkeiten und Herausforderungen von Lehrpersonen an allgemeinbildenden höheren Schulen in Bezug auf die standardisierte, schriftliche Reifeprüfung in Mathematik* [*Fears, difficulties and challenges of teachers at general secondary schools in relation to the standardized, written matriculation examination in mathematics*] [Master's thesis, Universität Wien].

Reisinger, M. (2020). *Mathematisches Bewusstsein in Zeiten eines technologieunterstützten Mathematik-Unterrichts* [*Mathematical awareness in times of technology-supported mathematics education*] [Seminar presentation]. FernUniversität Hagen.

Sattlberger, E. (2020). *Was wir leisten, wenn wir beurteile: Altes und Neues zum (leidigen) Thema Leistungsbeurteilung* [*What we achieve when we evaluate: Old and new on the (tedious) topic of performance evaluation*]. https://kphvie.ac.at/fileadmin/Dateien_KPH/Forschung_Entwicklung/Publikationen/KPH-Reihe/av-sattlberger-gesamtheft.pdf

Sattlberger, E., & Steinfeld, J. (2015). Die standardisierte schriftliche Reifeprüfung Mathematik (AHS)–Einsichten und Hintergrundinformationen [The standardized written mathematics school leaving examination (AHS)–Insights and background information]. *BIFIE Wien*. https://www.oemg.ac.at/DK/Didaktikhefte/2015%20Band%2048/VortragSattlbergerSteinfeld.pdf

Schnabl, C., & Kriegler-Kastelic, G. (2014). Studierfähigkeit auf dem Prüfstand. Kompetenz, Eignung und Begabung an der Schnittstelle von Schule und Hochschule [Ability to study put to the test. Competence, suitability and talent at the interface between school and university]. In S. Lin-Klitzing, D. Di Fuccia, & R. Stengl-Jörns (Eds.), *Gymnasium - Bildung - Gesellschaft: Abitur und Studierfähigkeit: Ein interdisziplinärer Dialog* [*High school-educatione-society: High school diploma and ability to study: An interdisciplinary dialogue*] (pp. 147-162). Klinkhardt.

Schulorganisationsgesetz. (1962/2023). *Bundesgesetz vom 25. Juli 1962 über die Schulorganisation (Schulorganisationsgesetz), RIS (1962 & rev. 04.03.2023)* [*Federal law of July 25, 1962, on school organization (School Organization Act), RIS (1962 & rev. March 4, 2023)*]. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009265

Singer, K. (2015). *Auswirkungen der Einführung einer standardisierten Reifeprüfung im Fach Mathematik in Österreich: Was passiert in der Praxis?* [*Effects of the introduction of a standardized matriculation examination in mathematics in Austria: What happens in practice?*] [Master's thesis, Karl-Franzens-Universität Graz].

Sorge, S., Petersen, S., & Neumann, K. (2016). Die Bedeutung der Studierfähigkeit für den Studienerfolg im 1. Semester in Physik [The importance of the ability to study for academic success in the first semester of physics]. *Zeitschrift Für Didaktik Der Naturwissenschaften* [*Journal for Didactics of Natural Sciences*], *22*(1), 165-180. https://doi.org/10.1007/s40573-016-0048-x

SRDP. (2021). Die standardisierte schriftliche Reifeprüfung in Mathematik (AHS) [The standardized written matriculation examination in mathematics (AHS)]. *Bundesministerium für Bildung, Wissenschaft und Forschung* [*Federal Ministry of Education, Science and Research*]. https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=4826&token=4574fed24b889f914a68a7411172dbce06459c69

SRDP. (2022). Mathematische Grundkompetenzen für die SRP in Mathematik (AHS) [Basic mathematical skills for the SRP in mathematics (AHS)]: Fundamental content areas for the core competencies in mathematics. *Bundesministerium für Bildung, Wissenschaft und Forschung* [*Federal Ministry of Education, Science and Research*]. https://www.matura.gv.at/downloads/download/mathematische-grundkompetenzen-fuer-die-srp-in-mathematik-ahs

Taschwer, K. (2018). Mathematiker: "Wir beobachten ein stetes Absinken der Kenntnisse" [Mathematician: "We are observing a constant decline in knowledge"]. *Der Standard*. https://www.derstandard.at/story/2000079538052/mathematiker-wir-beobachten-ein-stetes-absinken-der-kenntnisse

Thaler, B. (2021). Einfluss der schulischen Vorbildung auf den Studienerfolg. Abschluss und Verbleib im Studium bei fachnaher vs. nicht fachnaher schulischer Vorbildung [Influence of previous school education on academic success. Completion and retention in studies with subject-related vs. non-subject-related school education]. In A. Pausits, R. Aichinger, M. Unger, M. Fellner, & B. Thaler (Eds.), *Rigour and relevance: Hochschulforschung im Spannungsfeld zwischen Methodenstrenge und Praxisrelevanz* [*Rigor and relevance: University research in the area of tension between methodological rigor and practical relevance*] (pp. 179-200). Waxmann Verlag.

Thaller, B., & Juen-Kretschmer, C. (Eds.). (2016). *Beiträge zur Fachdidaktik: Projekt LEMMA: Zwischenbericht 2015* [*Contributions to subject didactics: LEMMA project: Interim report 2015*]. Praesens Verlag.

Thonhauser, J. (2020). Aufgaben als Katalysatoren von Lernprozessen [Tasks as catalysts of learning processes]. In U. Greiner, C. Wiesner, C. Schreiner, & F. Hofmann (Eds.), *Bildungsstandards: Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* [*Educational standards: competence orientation, task culture and quality development in the school system*] (pp. 464-481). Waxmann Verlag.

Tieben, N. (2019). Brückenkursteilnahme und Studienabbruch in Ingenieurwissenschaftlichen Studiengängen [Participation in bridging courses and dropping out of engineering courses]. *Zeitschrift Für Erziehungswissenschaft* [*Journal of Educational Science*], *22*(5), 1175-1202. https://doi.org/10.1007/s11618-019-00906-z

Tscholl, P. (2023). *Mathematical assessment for first-year students: Development, experience, and outlook at the University of Innsbruck.* European Society for Research in Mathematics Education.

Tscholl, P., Stampfer, F., & Hell, T. (Manuscript submitted for publication). About bridges and hurdles when entering university: Short-term effects of a mathematical bridging course concept and its self-assessment on academic success and dropout in Austria.

van den Berk, I., Stolz, K., Petersen, K., & Schultes, K. (2016). Einleitung [Introduction]. In I. van den Berk, K. Petersen, K. Schultes, & K. Stolz (Eds.), *Universitätskolleg-Schriften: Studierfähigkeit: Theoretische Erkenntnisse, empirische Befunde und praktische Perspektiven* [*University college writings: Study ability: Theoretical insights, empirical findings and practical perspectives*] (pp. 17-22).

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*.

Winkler, R. (2018). Zentralmatura in der Sackgasse? [Central high school diploma in a dead end?] *Internationale Mathematische Nachrichten* [*International Mathematical News*]*, 237*, 27-58.

Wößmann, L. (2002). Central exams improve educational performance: International evidence. *Kieler Diskussionsbeiträge* [*Kiel Discussion Contributions*]*, 397*, 1-45.

Wößmann, L. (2003). How central exams affect educational achievement: International evidence from TIMSS and TIMSS-repeat. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.387700

APPENDIX A

Table A1. Comparison of the experimental and control group on available control variables for sample case 1 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	22 (20 to 30)	20 (19 to 23)	2 (0 to 8)	0.066
Cohort (2020/21)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Cohort (2021/22)	5/17 (29.4%)	7/17 (41.2%)	1.65 (0.33 to 8.94)	0.721
Cohort (2022/23)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Gender (female)	6/17 (35.3%)	8/17 (47.1%)	1.61 (0.34 to 8.06)	0.728
First semester STEM (yes)	13/17 (76.5%)	15/17 (88.2%)	2.25 (0.27 to 28.77)	0.656
Last school grade (1)	5/17 (29.4%)	5/17 (29.4%)	1 (0.18 to 5.65)	1.000
Last school grade (2)	6/17 (35.3%)	6/17 (35.3%)	1 (0.20 to 5.11)	1.000
Last school grade (3)	3/17 (20.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Last school grade (4)	3/17 (17.6%)	2/17 (11.8%)	0.63 (0.05 to 6.39)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/17 (17.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Difficulties while solving mathematical test (lack of practice)	7/17 (41.2%)	7/17 (41.2%)	1 (0.21 to 4.82)	1.000
Difficulties while solving mathematical test (no technology use)	1/17 (5.9%)	2/17 (11.8%)	2.09 (0.10 to 133.48)	1.000
School type (AHS)	3/17 (17.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Self-concept score	4.17 (3.33 to 4.25)	4.33 (3.33 to 4.5)	-0.17 (-0.67 to 0.33)	0.333
Self-efficacy score	4.5 (4 to 4.67)	4.75 (4.25 to 4.75)	-0.25 (-0.50 to 0.00)	0.124
STEM (S)	7/17 (41.2%)	5/17 (29.4%)	0.6 (0.11 to 3.04)	0.721
STEM (T)	4/17 (23.5%)	3/17 (17.6%)	0.7 (0.09 to 5.07)	1.000
STEM (E)	2/17 (11.8%)	0/17 (0.0%)	0 (0.00 to 5.27)	0.485
STEM (M)	4/17 (23.5%)	9/17 (52.9%)	3.51 (0.69 to 21.30)	0.157

Table A2. Comparison of the experimental and control group on available control variables for sample case 2 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	21 (19.25 to 22)	20 (19.5 to 22.5)	0 (-2 to 2)	0.973
Cohort (2020/21)	5/11 (45.5%)	5/11 (45.5%)	1 (0.14 to 7.21)	1.000
Cohort (2021/22)	4/11 (36.4%)	5/11 (45.5%)	1.43 (0.20 to 11.17)	1.000
Cohort (2022/23)	2/11 (18.2%)	1/11 (9.1%)	0.47 (0.01 to 10.42)	1.000
Gender (female)	5/11 (45.5%)	3/11 (27.3%)	0.47 (0.05 to 3.59)	0.659
First semester STEM (yes)	9/11 (81.8%)	10/11 (90.9%)	2.14 (0.10 to 143.85)	1.000
Last school grade (1)	4/11 (36.4%)	3/11 (27.3%)	0.67 (0.07 to 5.59)	1.000
Last school grade (2)	4/11 (36.4%)	2/11 (18.2%)	0.41 (0.03 to 3.84)	0.635
Last school grade (3)	2/11 (18.2%)	5/11 (45.5%)	3.52 (0.40 to 48.84)	0.362
Last school grade (4)	1/11 (9.1%)	1/11 (9.1%)	1 (0.01 to 86.19)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/11 (27.3%)	3/11 (27.3%)	1 (0.10 to 9.94)	1.000
Difficulties while solving mathematical test (lack of practice)	4/11 (36.4%)	5/11 (45.5%)	1.43 (0.20 to 11.17)	1.000
Difficulties while solving mathematical test (no technology use)	1/11 (9.1%)	1/11 (9.1%)	1 (0.01 to 86.19)	1.000
School type (AHS)	1/11 (9.1%)	0/11 (0.0%)	0 (0.00 to 3.59)	1.000
Self-concept score	3.83 (3.42 to 4.17)	4 (3.33 to 4.50)	-0.17 (-0.67 to 0.67)	0.741
Self-efficacy score	4.5 (4.25 to 4.62)	4.25 (4.25 to 4.71)	0 (-0.50 to 0.25)	0.946
STEM (S)	4/11 (36.4%)	4/11 (36.4%)	1 (0.13 to 7.91)	1.000
STEM (T)	2/17 (18.2%)	2/17 (18.2%)	1 (0.06 to 16.69)	1.000
STEM (E)	2/17 (18.2%)	0/17 (0.0%)	0 (0.00 to 5.24)	0.476
STEM (M)	3/17 (27.3%)	5/17 (45.5%)	2.14 (0.28 to 19.77)	0.659

Table A3. Comparison of the experimental and control group on available control variables for sample case 3 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	22 (20 to 30)	20 (19 to 23)	2 (0 to 8)	0.089
Cohort (2020/21)	6/17 (35.3%)	7/17 (41.2%)	1.27 (0.26 to 6.42)	1.000
Cohort (2021/22)	5/17 (29.4%)	7/17 (41.2%)	1.65 (0.33 to 8.94)	0.721
Cohort (2022/23)	6/17 (35.3%)	3/17 (17.6%)	0.4 (0.05 to 2.43)	0.438
Gender (female)	6/17 (35.3%)	6/17 (35.3%)	1 (0.20 to 5.11)	1.000
First semester STEM (yes)	13/17 (76.5%)	15/17 (88.2%)	2.25 (0.27 to 28.77)	0.656
Last school grade (1)	5/17 (29.4%)	5/17 (29.4%)	1 (0.18 to 5.65)	1.000
Last school grade (2)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Last school grade (3)	3/17 (20.6%)	5/17 (29.4%)	1.91 (0.30 to 14.92)	0.688
Last school grade (4)	3/17 (17.6%)	2/17 (11.8%)	0.63 (0.05 to 6.39)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/17 (17.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Difficulties while solving mathematical test (lack of practice)	7/17 (41.2%)	8/17 (47.1%)	1.26 (0.27 to 6.07)	1.000
Difficulties while solving mathematical test (no technology use)	1/17 (5.9%)	2/17 (11.8%)	2.09 (0.10 to 133.48)	1.000
School type (AHS)	3/17 (17.6%)	3/17 (17.6%)	1 (0.11 to 8.82)	1.000
Self-concept score	4.17 (3.33 to 4.25)	4.17 (3.33 to 4.50)	-0.17 (-0.67 to 0.33)	0.545
Self-efficacy score	4.5 (4.00 to 4.67)	4.75 (4.25 to 5.00)	-0.25 (-0.50 to 0.00)	0.108
STEM (S)	7/17 (41.2%)	7/17 (41.2%)	1 (0.21 to 4.82)	1.000
STEM (T)	4/17 (23.5%)	3/17 (17.6%)	0.7 (0.09 to 5.07)	1.000
STEM (E)	2/17 (11.8%)	0/17 (0.0%)	0 (0.00 to 5.27)	0.485
STEM (M)	4/17 (23.5%)	7/17 (41.2%)	2.22 (0.42 to 13.48)	0.465

Table A4. Comparison of the experimental and control group on available control variables for sample case 4 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	22 (20 to 30)	20 (19 to 23)	2 (0 to 8)	0.080
Cohort (2020/21)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Cohort (2021/22)	5/17 (29.4%)	7/17 (41.2%)	1.65 (0.33 to 8.94)	0.721
Cohort (2022/23)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Gender (female)	6/17 (35.3%)	7/17 (41.2%)	1.27 (0.26 to 6.42)	1.000
First semester STEM (yes)	13/17 (76.5%)	13/17 (76.5%)	1 (0.15 to 6.65)	1.000
Last school grade (1)	5/17 (29.4%)	6/17 (35.3%)	1.3 (0.25 to 7.12)	1.000
Last school grade (2)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Last school grade (3)	3/17 (20.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Last school grade (4)	3/17 (17.6%)	2/17 (11.8%)	0.63 (0.05 to 6.39)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/17 (17.6%)	3/17 (17.6%)	1 (0.11 to 8.82)	1.000
Difficulties while solving mathematical test (lack of practice)	7/17 (41.2%)	8/17 (47.1%)	1.26 (0.27 to 6.07)	1.000
Difficulties while solving mathematical test (no technology use)	1/17 (5.9%)	2/17 (11.8%)	2.09 (0.10 to 133.48)	1.000
School type (AHS)	3/17 (17.6%)	5/17 (29.4%)	1.91 (0.30 to 14.92)	0.688
Self-concept score	4.17 (3.33 to 4.25)	4.17 (3.33 to 4.50)	-0.17 (-0.67 to 0.33)	0.628
Self-efficacy score	4.5 (4.00 to 4.67)	4.67 (4.25 to 4.75)	-0.25 (-0.50 to 0.00)	0.184
STEM (S)	7/17 (41.2%)	7/17 (41.2%)	1 (0.21 to 4.82)	1.000
STEM (T)	4/17 (23.5%)	3/17 (17.6%)	0.7 (0.09 to 5.07)	1.000
STEM (E)	2/17 (11.8%)	0/17 (0.0%)	0 (0.00 to 5.27)	0.485
STEM (M)	4/17 (23.5%)	7/17 (41.2%)	2.22 (0.42 to 13.48)	0.465

Table A5. Comparison of the experimental and control group on available control variables for sample case 5 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	22 (20 to 30)	20 (19 to 23)	2 (0 to 7)	0.086
Cohort (2020/21)	6/17 (35.3%)	7/17 (41.2%)	1.27 (0.26 to 6.42)	1.000
Cohort (2021/22)	5/17 (29.4%)	6/17 (35.3%)	1.3 (0.25 to 7.12)	1.000
Cohort (2022/23)	6/17 (35.3%)	4/17 (23.5%)	0.57 (0.09 to 3.17)	0.708
Gender (female)	6/17 (35.3%)	8/17 (47.1%)	1.61 (0.34 to 8.06)	0.728
First semester STEM (yes)	13/17 (76.5%)	15/17 (88.2%)	2.25 (0.27 to 28.77)	0.656
Last school grade (1)	5/17 (29.4%)	5/17 (29.4%)	1 (0.18 to 5.65)	1.000
Last school grade (2)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Last school grade (3)	3/17 (20.6%)	5/17 (29.4%)	1.91 (0.30 to 14.92)	0.688
Last school grade (4)	3/17 (17.6%)	2/17 (11.8%)	0.63 (0.05 to 6.39)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/17 (17.6%)	3/17 (17.6%)	1 (0.11 to 8.82)	1.000
Difficulties while solving mathematical test (lack of practice)	7/17 (41.2%)	7/17 (41.2%)	1 (0.21 to 4.82)	1.000
Difficulties while solving mathematical test (no technology use)	1/17 (5.9%)	2/17 (11.8%)	2.09 (0.10 to 133.48)	1.000
School type (AHS)	3/17 (17.6%)	3/17 (17.6%)	1 (0.11 to 8.82)	1.000
Self-concept score	4.17 (3.33 to 4.25)	4.17 (3.33 to 4.50)	-0.17 (-0.67 to 0.33)	0.545
Self-efficacy score	4.5 (4.00 to 4.67)	4.75 (4.25 to 5.00)	-0.25 (-0.50 to 0.00)	0.108
STEM (S)	7/17 (41.2%)	6/17 (35.3%)	0.78 (0.16 to 3.84)	1.000
STEM (T)	4/17 (23.5%)	2/17 (11.8%)	0.44 (0.03 to 3.7)	0.656
STEM (E)	2/17 (11.8%)	0/17 (0.0%)	0 (0.00 to 5.27)	0.485
STEM (M)	4/17 (23.5%)	9/17 (52.9%)	3.51 (0.69 to 21.3)	0.157

Table A6. Comparison of the experimental and control group on available control variables for sample case 6 (for continuous variables, comparison was conducted using two-sided two-sample Wilcoxon rank sum test [estimated median difference as estimate] & for categorical variables using Fisher's exact test [odds ratio as estimate])

	Control group (n=17)	Experimental group (n=17)	Estimate with 95% confidence interval	p-value
Age	22 (20 to 30)	20 (19 to 23)	2 (0 to 8)	0.074
Cohort (2020/21)	6/17 (35.3%)	6/17 (35.3%)	1 (0.20 to 5.11)	1.000
Cohort (2021/22)	5/17 (29.4%)	7/17 (41.2%)	1.65 (0.33 to 8.94)	0.721
Cohort (2022/23)	6/17 (35.3%)	4/17 (23.5%)	0.57 (0.09 to 3.17)	0.708
Gender (female)	6/17 (35.3%)	7/17 (41.2%)	1.27 (0.26 to 6.42)	1.000
First semester STEM (yes)	13/17 (76.5%)	15/17 (88.2%)	2.25 (0.27 to 28.77)	0.656
Last school grade (1)	5/17 (29.4%)	5/17 (29.4%)	1 (0.18 to 5.65)	1.000
Last school grade (2)	6/17 (35.3%)	5/17 (29.4%)	0.77 (0.14 to 4.05)	1.000
Last school grade (3)	3/17 (20.6%)	5/17 (29.4%)	1.91 (0.30 to 14.92)	0.688
Last school grade (4)	3/17 (17.6%)	2/17 (11.8%)	0.63 (0.05 to 6.39)	1.000
Difficulties while solving mathematical test (lack of motivation/concentration)	3/17 (17.6%)	4/17 (23.5%)	1.42 (0.20 to 11.62)	1.000
Difficulties while solving mathematical test (lack of practice)	7/17 (41.2%)	7/17 (41.2%)	1 (0.21 to 4.82)	1.000
Difficulties while solving mathematical test (no technology use)	1/17 (5.9%)	2/17 (11.8%)	2.09 (0.10 to 133.48)	1.000
School type (AHS)	3/17 (17.6%)	3/17 (17.6%)	1 (0.11 to 8.82)	1.000
Self-concept score	4.17 (3.33 to 4.25)	4.17 (3.33 to 4.50)	-0.17 (-0.67 to 0.33)	0.436
Self-efficacy score	4.5 (4.00 to 4.67)	4.75 (4.25 to 5.00)	-0.25 (-0.50 to 0.00)	0.108
STEM (S)	7/17 (41.2%)	6/17 (35.3%)	0.78 (0.16 to 3.84)	1.000
STEM (T)	4/17 (23.5%)	3/17 (17.6%)	0.7 (0.09 to 5.07)	1.000
STEM (E)	2/17 (11.8%)	0/17 (0.0%)	0 (0.00 to 5.27)	0.485
STEM (M)	4/17 (23.5%)	8/17 (47.1%)	2.8 (0.54 to 16.9)	0.282

Table A7. Composition of sample cases (nine people from the experimental group are in each sample case)

		Sample cases					
	Subject ID	1	2	3	4	5	6
Control group	c1	x		x	x	x	x
	c2	x		x	x	x	x
	c3	x		x	x	x	x
	c4	x	x	x	x	x	x
	c5	x	x	x	x	x	x
	c6	x	x	x	x	x	x
	c7	x	x	x	x	x	x
	c8	x	x	x	x	x	x
	c9	x	x	x	x	x	x
	c10	x	x	x	x	x	x
	c11	x		x	x	x	x
	c12	x	x	x	x	x	x
	c13	x	x	x	x	x	x
	c14	x		x	x	x	x
	c15	x	x	x	x	x	x
	c16	x		x	x	x	x
	c17	x	x	x	x	x	x
Experimental group	e1	x	x	x	x	x	x
	e2	x	x	x	x	x	x
	e3	x	x	x	x	x	x
	e4	x	x	x	x	x	x
	e5	x	x	x	x	x	x
	e6	x	x	x	x	x	x
	e7	x	x	x	x	x	x
	e8	x	x	x	x	x	x
	e9	x	x	x	x	x	x
	e10	x		x	x	x	x
	e11	x		x	x	x	x
	e12	x		x	x	x	x
	e13	x		x	x	x	x
	e14	x	x	x	x		x
	e15		x	x		x	x
	e16	x			x		x
	e17	x		x		x	
	e18	x				x	x
	e19			x			
	e20					x	
	e21				x		
	e22				x		