Closing the gender STEM gap
A randomized-controlled trial in elementary schools

Kerstin Grosch*[1], Simone Haeckl †[2], and Martin G. Kocher‡[3]

[1]WU Vienna University of Economics and Business & Institute for Advanced Studies Vienna (IHS)
[2]University of Stavanger
[3]University of Vienna & University of Gothenburg

March 7, 2024

Abstract

We examine whether a digital web application for elementary-school children can increase children's interest in STEM with a specific focus on narrowing the gender gap. Coupling a randomized-controlled trial with experimental lab and survey data, we analyze the effect of the digital intervention and shed light on potential behavioral mechanisms. We demonstrate that girls have a lower overall interest in STEM than boys. Our treatment increases girls' interest in STEM and, consequently narrowing the gender gap by approximately 20%. This outcome establishes the malleability of STEM interests. Our findings further suggest that an easy-to-implement digital intervention has the potential to foster gender equality for young children and can potentially contribute to a reduction of gender inequalities in the labor market in terms of occupational sorting and the gender wage gap later in life.

JEL Classification numbers: C93, D91, I24, J16, J24
Keywords: STEM, digital intervention, gender equality, field experiment

*Vienna University of Economics and Business (WU), Welthandelsplatz 1, 1020 Vienna, Austria; *email: kerstin.grosch@wu.ac.at*

†University of Stavanger, Business School, Stavanger, Norway; *email: simone.haeckl-schermer@uis.no*

‡University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; *email: martin.kocher@univie.ac.at*

Special thanks to the Ministry of Education, the Education Directorates in Vienna and Upper Austria, Maria Theresia Niss, FehrAdvice & Partners AG, the Content Performance Group GmbH, and to our team of research assistants who supported the fieldwork. We thank audiences at the Institute for Advanced Studies, the Max-Planck-Institute in Bonn, the Vienna University of Economics and Business, the CESifo Research Network Area Meeting Economics of Education, and Christina Felfe, Ben Greiner, Marcela Ibanez, Rupert Sausgruber, and Matthias Sutter for very helpful comments. All errors are ours. This research was funded in part by the Austrian Science Fund (FWF) [T 1263-G]. Moreover, the study was supported by the Institute for Advanced Studies, the B&C Privatstiftung, the Federation of Austrian Industries (IV), and the University of Stavanger. The project received IRB approval from the ethics committee at the Institute for Advanced Studies (IHS) in Vienna on October 18th, 2019, CASE 02 MINT 2020000 IA and is pre-registered at the AEA RCT Registry before the intervention and data collection (AEARCTR-0005014).

1 Introduction

Women are less likely than men to specialize in STEM fields, i.e., in science, technology, engineering, and mathematics, in many Western countries (Ginther and Kahn, 2004; Bertrand et al., 2010). We call this phenomenon the gender STEM gap. Given that these jobs provide, on average, higher wages, hold relatively good chances for permanent employment, increase the likelihood of taking up leadership positions, and are socially relevant to meet global challenges (Webber, 2014; Cedefop, 2015; Joensen and Nielsen, 2016; Kahn and Ginther, 2018; ILO, 2019; UNESCO, 2021)[1], it is important to learn more about potential causes for the gender STEM gap and to develop effective interventions to close it.

The understanding of the underlying reasons for women's underrepresentation in STEM fields remains surprisingly limited. The most extensively studied aspects revolve around STEM abilities, particularly math performance, where gender differences are commonly observed in both average performance and variance (e.g., Hyde et al., 2008; Fryer Jr and Levitt, 2010; Nollenberger et al., 2016). However, these gender differences in ability explain the gender STEM gap only to a limited extent (Hyde et al., 2008; Zafar, 2013; Wiswall and Zafar, 2015; Kahn and Ginther, 2018; Jiang, 2021; Saltiel, 2023). Instead, recent research indicates that personal tastes and individual attributes play a more significant role (e.g., Wiswall and Zafar, 2015; Saltiel, 2023). Our study builds on this understanding and shifts the focus from abilities to interests, i.e., the natural inclination towards STEM and non-STEM fields. More precisely, we investigate whether STEM interests are malleable through a digital intervention. Consequently, we designed a digital treatment intervention, a web-based application, prioritizing creating interest in STEM over raising their abilities with the aim of increasing girls' inherent inclination towards STEM disciplines.

To investigate the intervention's impact and investigate behavioral mechanisms for the gender STEM interest gap, we employ a large-scale randomized-controlled trial (RCT) in elementary schools in Austria combined with individual survey and experimental measures before and after the intervention. The treatment app presents STEM fields in an engaging way and by addressing the underlying behavioral mechanisms that could interfere with the development of interest in STEM. The treatment app presents both male and female STEM professionals, such as engineers and programmers, on fantasy planets. Accompanied by the professionals, the children playfully learn more about various societal challenges, such as threats from climate change and public health, and how STEM skills can contribute to combating them. The app comprises exercises, videos, and texts and informs children about STEM-related content in general. To enhance the app's effectiveness, we identified four potential behavioral mechanisms contributing to the

[1]The higher wages in STEM professions can be explained by relatively high demand in these jobs coupled with the specific skill set that job candidates must acquire to work in these fields (Deming and Noray, 2020). Obviously, an intervention such as ours will not shift the equilibrium wages through a much higher general supply in STEM professionals or lead to a change of wages in the profession due to an increased share of the female workforce (Levanon et al., 2009).

development of STEM interests. First, children may be more likely to associate STEM with boys than with girls. Consequently, girls with more pronounced stereotypical beliefs concerning STEM and gender may demonstrate less interest in STEM. Second, a growth mindset, i.e., believing that you can increase your performance by practicing and using the appropriate learning techniques, may be important for establishing an interest in STEM (Kahn and Ginther, 2018; Bettinger et al., 2018; Alan et al., 2019). Third, Buser et al. (2017) show that more pronounced competitive preferences can help explain the interest in STEM and, ultimately, the decision for a STEM career.[2] Fourth, confidence, the belief in one's abilities, may be relevant for building up an interest in STEM (Carlana, 2019; Saltiel, 2023). These behavioral mechanisms are systematically addressed in the treatment app by, e.g., using child-friendly tutorials with visualizations on synapse creation in the brain when learning to promote a growth mindset, introducing female STEM role models in videos to overcome stereotypical beliefs, and awarding STEM badges as positive feedback for boosting children's confidence. We collect ex-ante and ex-post measurements to increase our understanding of the role of these individual characteristics for STEM interest.

We recruited 39 elementary schools with 60 classes in the third grade in Vienna (an urban area) and Upper Austria (a predominantly rural area). We randomly assigned about half of the schools to the treatment group and the other half to the control group. The treatment group used the treatment app *with* a STEM focus as described above. The control group used a "traditional" learning app without a STEM focus (control app) which retained critical features such as exposure to a technical device consistent, thereby minimizing potential confounding variables when identifying treatment effects. We implemented the study as an out-of-school intervention and asked children and their parents to engage with the app each weekday for about ten minutes for four weeks across control and treatment groups. We collected data in class using incentivized decisions and survey measures before (baseline) and after the intervention.

Our RCT focuses on elementary schoolchildren for various reasons. First, aptitudes toward particular educational fields seem likely to develop early in life (Tai et al., 2006; DeWitt et al., 2013). That is, boys and girls, start to differ in crucial determinants for educational decisions such as preferences and beliefs already at a young age (e.g., Dahlbom et al., 2011; Buser et al., 2014; Bian et al., 2017). Second, interests will guide children's skill investments, which are important inputs into their education production functions, shaping later life choices and outcomes (Heckman, 2006; List et al., 2018, 2021). Relevant education choices such as track and school choices are made as early as after elementary school in Austria at an age below or around ten

[2]Competitive preferences are, as in this study, commonly proxied by the decision to enter a competition. Recent findings by Gillen et al. (2019) and Van Veldhuizen (2022) started a discussion about the importance of competitive preferences in addition to risk preferences and overconfidence. For our study, we deliberately abstain from disentangling different behavioral foundations of the concept of competitiveness because of the existing evidence on the positive association of competitiveness and STEM interest in other studies and time constraints during the study's implementation that would not have allowed for measuring all relevant aspects. In order to be precise and to distinguish properly between concepts, we refer to our measure as either "competitiveness" or "competitive aptitude" instead of "competitive preferences throughout the paper."

years and might be important in determining a child's preparedness to pursue a STEM career later in life (Delaney and Devereux, 2019; Card and Payne, 2021). Third, interventions may be more effective at an early age when particular interests and preferences are still malleable (e.g., Heckman et al., 2010; Cappelen et al., 2020; Kosse et al., 2020). Due to logistical constraints and resource limitations, we needed to gather individual-level data at the classroom rather than at the individual level which requires a certain degree of cognitive development and a sufficient attention span (e.g., Grosch et al., 2023; List et al., 2023). As a result, we did not include first or second graders in our study.

We analyze data from 962 children. Our data confirm that the gender gap in STEM interest emerges already in elementary school, with girls having a significantly lower interest in STEM than boys. Our treatment app increases girls' interest in STEM significantly, and, as a consequence, narrows the gender gap in STEM interest between boys and girls by about 20%. In a mediation analysis, we show that our measure of STEM confidence explains the majority of the treatment effect on girls' STEM interests. The treatment also increases girls' competitiveness by seven percentage points. However, we do not find a mediating effect from competitiveness on STEM interest.

Our study makes several contributions to the literature on the gender STEM gap. First of all, our study is the first to investigate the malleability of STEM interests. For this, we developed a novel measure of STEM interest that is appropriate for our study, taking into account the age of our young participants and the country-specific context. We demonstrate the validity of this measure with the common approach in economics of correlating survey measures with real-life or incentivized choices (e.g., Dohmen et al., 2011; Sutter et al., 2013; Enke et al., 2022; Falk et al., 2023). The measure enables us to demonstrate that the gender gap in STEM interest is evident as early as nine years old and, most importantly, that STEM interests are malleable. This finding suggests that behaviorally-informed interventions can be an important tool for closing the gender STEM gap. By that, our study contributes to the literature on behavioral early-childhood interventions aimed at improving later labor-market outcomes. While the bulk of this research has focused on cognitive skills, more recently, the importance of non-cognitive skills and personality traits for labor-market success has received more attention (e.g., Heckman, 2000; Currie, 2001; Almlund et al., 2011; Kautz et al., 2014; Cappelen et al., 2020; Kosse et al., 2020).[3] For example, Alan et al. (2019) and Alan and Ertac (2018a) succeed in increasing children's grit and patience through school interventions. We contribute to this literature by demonstrating, in the context of a large-scale intervention study, how girl's interest in STEM, which is strongly correlated with favorable labor-market outcomes, can be fostered.

Early-childhood interventions aiming at fostering gender equality in educational and labor

[3]Many of these studies put an emphasis on children from low socio-economic status or underrepresented families (see, for example, Heckman et al., 2010; Cook et al., 2014; Heller, 2014; Oreopoulos et al., 2017; Kosse et al., 2020; Cohodes et al., 2022). Other studies with children show the effectiveness of interventions at improving later labor-market outcomes by fostering traits such as a growth mindset, patience, and competitiveness, without a focus on a specific target group (e.g., Alan and Ertac, 2018a,b; Sorrenti et al., 2020; Rege et al., 2021).

market outcomes such as ours are scant. An exception is research on gender differences in competitiveness. We know from numerous studies that women are less inclined to compete than men, and studies demonstrate a link between higher competitiveness and more favorable educational and labor-market choices/outcomes (Buser et al., 2014; Flory et al., 2015; Buser et al., 2021). Alan and Ertac (2018b) and Hermes et al. (2021) demonstrate in randomized controlled trials that their grit and feedback interventions, respectively, reduce the gender gap in competitiveness. We contribute to this literature by developing an intervention that aims at attenuating the gender STEM gap. Ultimately, such a reduction in the gender STEM gap should improve women's labor-market outcomes.

Other research related to our study deals with online education apps that provide an easily scalable, cost-effective way of learning at an individual pace with individual feedback, holding the potential to improve educational outcomes including STEM skills (Mayo, 2009; Berkowitz et al., 2015; Escueta et al., 2020). Berkowitz et al. (2015) demonstrate that an app for first graders and their parents used out-of-school can reduce parents' math anxiety and, consequently, improve children's math ability. We add to the literature by demonstrating that an easy-to-use and not very costly app applied out-of-school can increase girls' STEM interest.

There is a large scholarly literature on potential underlying mechanisms for the gender STEM gap (for an overview see McNally, 2020). The majority of these studies focus on adolescents, their STEM abilities, and educational decisions later in life (e.g., Buser et al., 2014, 2017; Delaney and Devereux, 2019; Card and Payne, 2021), e.g., the influence of the gender composition in high school classes on the propensity to become a STEM professional (e.g., Brenøe and Zölitz, 2020), the role of prior course choices on STEM degree selection (Delaney and Devereux, 2019; Card and Payne, 2021), or the effects of teachers' stereotypical bias on girls' math performance and confidence (Carlana, 2019). We add to this literature by examining individual attributes and their impact on children's interest in STEM at a young age.

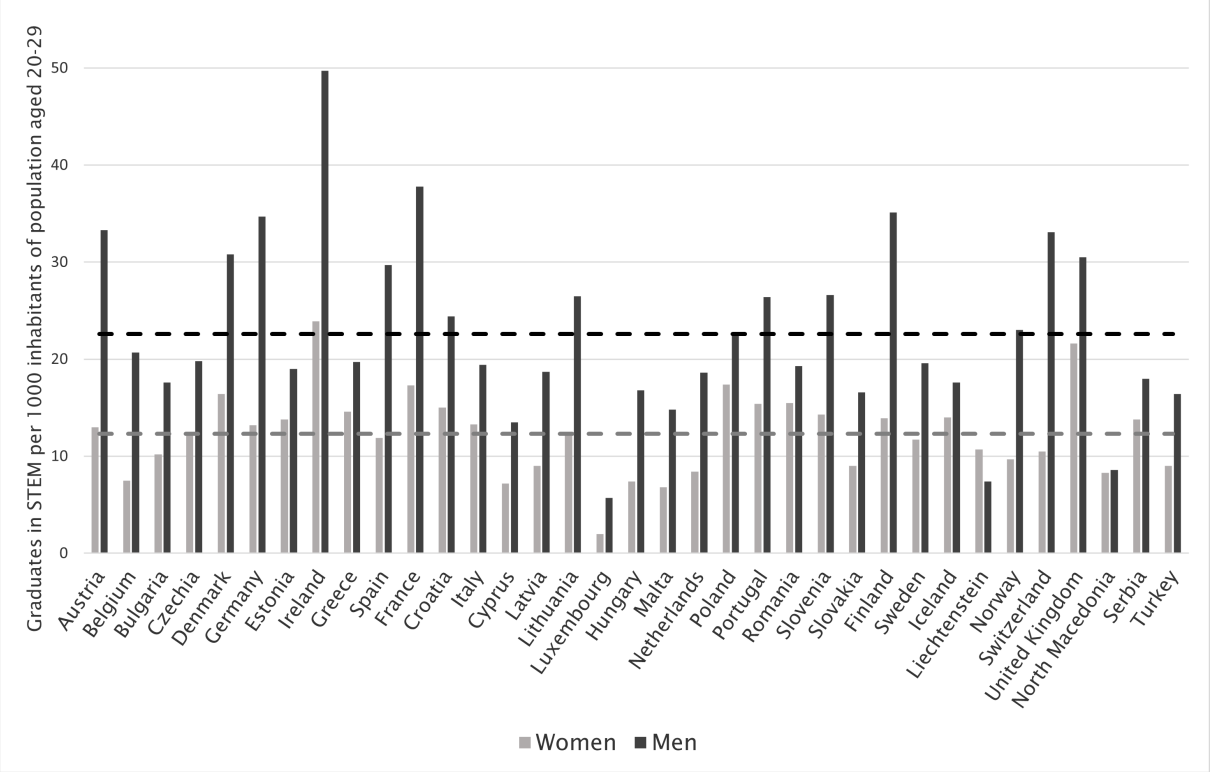
The remainder of this paper is organized as follows. In Section 2, we describe the institutional background of our research sites. In Section 3, we present details of the intervention, discuss our hypotheses, explain the study design, and provide internal balancing tests. Section 4 presents results first on baseline gender differences, then on general treatment effects, and it eventually examines the role of underlying behavioral mechanisms on STEM interest. In Section 5, we discuss our results, potential limitations, and policy implications. We conclude in Section 6.

2 Institutional background and country context

In Austria, the unadjusted gender pay gap was 18.9% in 2020, which is higher than the EU average of 13% (Eurostat, 2018). A large proportion of the gender pay gap can be explained by differences in payment structures of economic sectors and gender differences in occupational choice (Brown and Corcoran, 1997; Blau and Kahn, 2000, 2017). In Austria, men graduate about three times more often from STEM study programs, i.e., in science, mathematics, computing,

engineering, manufacturing, and construction, than women (Eurostat, 2019). The gender difference in STEM graduates in Austria is relatively large (20.3 percentage points) compared to the average in European countries (10.3 percentage points). We illustrate the gender differences in STEM graduates across European countries in Figure 1.

FIGURE 1
Share of female and male STEM graduates across European Countries
(source: EUROSTAT 2019; own illustration)



Generally, after four years of elementary school, children in Austria move to lower secondary school at the age of ten years. They can choose either a general secondary school (middle school) or an academic secondary school (high school), partly depending on their performance in elementary school. Lower secondary schools can have specific specializations such as languages, music education, or STEM. The specialization is implemented either through additional subjects offered in school or through an increased number of class hours in some of the subjects. After four years in a lower secondary school, at age 14 teenagers can either choose to attend a school preparing them for vocational training, stay in the academic secondary school, or attend a college for higher vocational training.[4] While female students are over-represented in colleges with a focus on business administration or health care, the share of female students choosing to attend a school with a technical specialization is 26% and thus rather low (Statistik Austria, 2019). Focusing on the three most popular options that teenagers choose for their vocational training, women are most likely to become saleswomen, office managers, or hairstylists, while

[4]Our description simplifies the option set; it focuses on the most common options.

the top three vocational tracks for boys are metal technology, electrical engineering, and automobile technology (Wirtschaftskammer Oesterreich, 2020). An analogous picture regarding gender differences emerges when focusing on tertiary education.

3 Study design

We preregistered the study design and the analysis at the AEA RCT Registry. For the write-up of the paper, we make very few adaptations to the preregistration that we did not foresee at the time of preregistration, following Banerjee et al. (2020). We preserve transparency by indicating the few changes (some variable definitions) explicitly and by providing an Online Appendix where we execute all the analyses as defined in the preregistration.[5]

3.1 Experimental groups

The core of our experiment consists of a treatment group that uses the treatment app and a control group that uses the control app. We implemented a sub-treatment informing parents of the usefulness of STEM skills for children using a cross-over design. Below, we describe our treatment in detail.

Treatment app

The treatment app is called "Robitopia" and is designed for an average usage time of ten minutes per school day over four weeks (see the Online Appendix for visualizations of the treatment app). It features an animated journey for children with a spaceship through a fantasy universe. Research has shown that children have biased beliefs about STEM professionals and their contributions to societal challenges. Such biased beliefs could constitute an impediment to the development of STEM interest, particularly in girls (Shin et al., 2019). The treatment app counteracts such biased beliefs, by letting children visit four different planets that face different societal challenges each week, relating the mastering of these challenges to STEM skills. Children use insights from STEM provided by STEM professionals to solve the weekly challenges step by step each day. In the treatment app, there is, for example, a planet with clean air, and children explore the technologies used and the ways of living that inhabitants follow to maintain air quality. Within the story, the children watch videos, read texts, play games, and meet avatars of natural scientists as well as other STEM professionals. While playing, children earn stars when they read, watch tutorials, and solve exercises that engage them playfully. Stars are also a game currency and can be used to pimp the spaceship. STEM professionals on the app are from different disciplines and education levels, ranging from engineers and programmers to technical assistants.

[5]The preregistration can be found at: https://doi.org/10.1257/rct.5014-1.1 or in the Online Appendix of this paper provided on our websites: https://sites.google.com/view/kerstin-grosch/research and https://sites.google.com/view/simonehaeckl

The treatment app also addresses the aforementioned behavioral mechanisms. To foster the children's growth mindset, the app contains a tutorial about how the brain cells connect each time they learn something and work hard to solve tasks. To increase confidence and competitiveness, we let children guess their performance after they have worked on one of the app's exercises and reward them for correct guesses with the "star" (game) currency. Moreover, children earn badges, e.g., a math or a science badge, when they deal with content in the app. To decrease stereotypical thinking about STEM ability, the treatment app shows mixed-gender (and mixed-ethnic) teams and videos in which young female STEM professionals introduce themselves, serving as potential role models (à la Bettinger and Long, 2005; Porter and Serra, 2020; Breda et al., 2023) and counteracting existing stereotypes about STEM professionals (Miller et al., 2018).

Control app

In the control group, children used an established learning app called "Anton," which is publicly available and supported by the European Union. This app primarily focuses on language (German) and mathematics exercises for various school levels. The purpose of the control group was to mitigate potential confounds to the identification of treatment effects. More specifically, we identified four potential confounding factors ex-ante. First, we wanted to keep the exposure to technology similar between the treatment and the control group as exposure to technology could per se affect interest in STEM. Second, we wanted to keep the extra time children spend on school-related activities similar in the two groups, and, third, the choice of the control app allowed us to keep teachers and children blind to the treatment. Lastly, we wanted to be able to identify children that are more likely to use an app that is promoted by the teacher to get better insights into the selection of app usage. "Anton" serves us very well on all four counts.

To ensure full comparability between the control and treatment conditions in terms of usage time, we created a group for each class. Each day children logged into the control app, they saw content to work on for about ten minutes which is similar to the daily engagement time for children in the treatment app. We pinned varying German and Math exercises that were suitable to their grade on the online board of the app each day. Hence, similar to the treatment app, children also partly worked on math-related tasks in the control app. However, the control app is not geared towards STEM, i.e., children worked on math tasks that were not associated with STEM content as in the treatment app. This way, the control app does not aim to overcome the stereotypical biases children might have had about math, e.g., that math is not needed to contribute to societal challenges. Moreover, contrary to the treatment app, the control app does not address other underlying behavioral mechanisms such as confidence but poses plain knowledge questions without further context. The fact that the control app also includes math questions makes finding a treatment effect harder.

Sub-treatment: brochures

In a sub-treatment, we handed out information brochures for parents (see the Online Appendix for the original version in German). These brochures are designed to inform parents about the importance of STEM and to advise ways to increase children's interest in science and technology. Importantly, we only handed out brochures when the main intervention period had ended to avoid spillover effects (see the experimental schedule in Section 3.4). We present the results of the brochure intervention in Appendix D and provide evidence that there is no interaction between the interventions in Table A.4 in the Appendix.

3.2 Hypotheses

As described above, the treatment app playfully introduces children to STEM topics and takes them on an individual journey (e.g., Mayo, 2009). The app counteracts biased beliefs about STEM fields, e.g., that STEM professionals work on abstract problems rather than contributing to finding solutions to societal challenges. Moreover, children actively engage with STEM-related content in different learning environments, i.e., tutorials, videos, and hands-on exercises. In addition, the app addresses behavioral mechanisms, including stereotypical thinking, growth mindset, competitiveness, and confidence, all of which may influence children's interest in STEM. Consequently, we anticipate that utilizing the treatment app will lead to an increase in children's interest in STEM.

Hypothesis 1:
The treatment app increases children's interest in STEM compared to the control app.

Fewer women than men specialize in STEM professionally (Blau and Kahn, 2000; Blau et al., 2013). Moreover, research has shown that adolescent girls are less interested in STEM than boys (e.g., Buser et al., 2017; Carlana, 2019). Furthermore, evidence demonstrates that preferences, skills, and tastes of girls and boys differ already in elementary school (e.g., Fehr et al., 2008; Sutter et al., 2019; Brocas and Carrillo, 2021). Bringing these strands of the literature together, we expect that, on average, girls are less interested in STEM fields than boys already in elementary school. Assuming that girls exhibit less pronounced baseline preferences for STEM, there is more room for girls to increase their STEM interests than for boys. Hence, we expect that girls respond more strongly to the treatment app than boys.

Hypothesis 2:
On average, girls exhibit lower levels of interest in STEM fields than boys. The treatment app leads to a higher increase in STEM interest for girls than for boys.

We identified four behavioral mechanisms potentially affecting girls' interest in STEM. First, children may hold stereotypical beliefs of boys' and girls' performance in a stereotypical male context such as math compared to a stereotypical female context such as language (Reuben

et al., 2014; Kollmayer et al., 2018). In general, STEM fields are associated more with men than with women (e.g., Shapiro and Williams, 2012). This stereotype forms as early as in elementary school (Cvencek et al., 2011; Miller et al., 2018). When girls do not envision themselves in STEM-related fields, they are deterred from entering by the potentially wrong perception that STEM is a "men's field" where girls do not belong (Shapiro and Williams, 2012).[6] Hence, we expect that girls with stronger stereotypical beliefs are less interested in STEM.

Second, we examine the effect of the children's way of thinking about the nature of talents, i.e., whether they are fixed or can develop in response to a dedicated effort. A "growth mindset" indicates the latter way of thinking. There is evidence that children who acquire a growth mindset are more likely to enroll in advanced mathematics courses and demonstrate higher levels of perseverance and performance in mathematics (e.g., Blackwell et al., 2007; Bettinger et al., 2018; Alan et al., 2019; Yeager et al., 2019). Mathematics is an essential aspect of most STEM subjects. Hence, we expect that having more of a growth mindset is positively associated with STEM interests.

Third, STEM subjects are often perceived as more prestigious and competitive than other fields such as the social sciences (Buser et al., 2014). In line with this view, it has been found that more pronounced competitiveness is associated with STEM interest and, ultimately, the decision for a science focus later in school (Buser et al., 2014, 2017). In addition, there is evidence that gender differences in the willingness to compete exist (e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2010; Tungodden and Willén, 2023) and emerge early in life (e.g., Dreber et al., 2014; Sutter and Glätzle-Rützler, 2015). Therefore, we expect that, on average, girls are less competitive than boys. Because of the lower baseline level of girls' competitiveness, the treatment app might be particularly successful in promoting competitiveness in girls.

Fourth, self-confidence in STEM-relevant skills is important for choosing a STEM track. An empirical study by Saltiel (2023) demonstrates that self confidence in math is an important predictor for STEM enrolment in college.[7] A study by Murphy and Weinhardt (2020) demonstrate that the academic rank in class in primary school has lasting effects, particularly on boys' confidence in their abilities, resulting in a larger number of math courses chosen at the end of secondary school. Hence, we expect that STEM confidence and STEM interest are positively related. In addition, studies have shown that girls are less confident in STEM-related subjects and their academic performance than boys (e.g., Dahlbom et al., 2011; Shi, 2018; Exley and

[6] To reduce stereotypes, children can be exposed to counter-stereotypical role-models (Olsson and Martiny, 2018). Studies empirically investigating the effect of those role models range from correlational evidence, controlling for the employment status of the mother (Eagly et al., 2000), exploiting exogenous variation in teachers' gender (e.g., Eble and Hu, 2020; Dee, 2005, 2007) to randomized control trials exposing participants randomly to certain role models (e.g., Porter and Serra, 2020; Breda et al., 2023)). The latter studies are closely related to this paper as they study the effect of short-term interventions with experimental methods.

[7] Saltiel (2023) uses a confidence measure called "math self-efficacy," based on students' responses to statements like "Confident I can do an excellent job on my math tests." Our measure is similar, as we also ask students about their confidence in specific STEM (and non-STEM) areas. However, our approach expands to cover STEM fields with substantial gender gaps and is not confined to math.

Kessler, 2022; Saltiel, 2023), suggesting that the effect of the app is more pronounced for girls than for boys.

Hypothesis 3: Behavioral mechanisms for an increase in STEM interest
(a) Stereotypical thinking: Girls are more interested in STEM with a decrease in the strength of their stereotypical beliefs.
(b) Growth mindset: Children are more interested in STEM with an increase in the level of their growth mindsets.
(c) Competitiveness: Children that are more competitive exhibit more interest in STEM. Since girls are, on average, less competitive than boys, the effect of the app is more pronounced for girls than for boys.
(d) Confidence: Children are more interested in STEM with increasing STEM confidence. Since girls are, on average, less confident than boys, the effect of the app is more pronounced for girls than for boys.

The treatment app intends to address the potential behavioral mechanisms that foster interest in STEM, given the existing literature. We expect these *indirect* mechanisms to partly explain the treatment effect. The mechanism variables will not explain the entire treatment effect (Hypothesis 1) due to unobserved mediators.

Hypothesis 4:
The behavioral mechanism variables partly explain the impact of the treatment app on STEM interest.

3.3 Measures[8]

Outcome measures

To proxy interest in STEM for elementary school children, we have developed two outcome variables.[9] Our first measure uses the description of different jobs. We describe six different occupations, three of which are STEM jobs (programming, engineering/technical worker, mathematician), and three are non-STEM jobs (social/health worker, journalist/writer/translator,

[8]The complete instructions for children's decisions in the experiment as well as minor changes to the preregistration can be found in the Online Appendix. As a robustness check, we elicited further measures for explicit stereotypes, implicit stereotypes, and confidence. These measures are discussed in Section C in the Appendix.

[9]We refrained from using existing measures in the educational literature that have been employed for older children or in other countries to proxy science or STEM interest (e.g., DeWitt et al., 2013; Kier et al., 2014) for different reasons. First, they use terms such as "science" in their survey questions that are not understood by elementary school children in the Austrian context since there is no such school subject or related subjects at that age, yet. Second, our study uses measures that reflect the variety of STEM occupations and are not restricted to science jobs only. Third, children in elementary school have a very limited attention span and might struggle with long scales in group data collection. Therefore, we refrained from measures with extended time requirements such as the scales mentioned above that contain more than 40 items.

arts and humanities). The descriptions use simple language and provide children with examples of what people do in these jobs. We based our selection of STEM jobs on the classification by the European Commission (European Commission, 2015). We chose the "core STEM jobs" that exclude jobs with no, or less pronounced, occupational gender gaps. The three non-STEM jobs are occupations in which women are typically over-represented (e.g., European Institute for Gender Equality, 2017).

After each job description, children indicated how much interest they had in the job on a Likert scale from one to five. To obtain an index outcome variable for our analysis, we sum up the scores for the STEM jobs and divide this sum by the sum of scores of both STEM and non-STEM jobs. More formally, let l_i be the indicated Likert scale value, and the integer i can take up numbers in the range of $[1, 6]$ for the six occupations; with $[1, 3]$ for the three STEM occupations and $[4, 6]$ for the three non-STEM occupations. The STEM interest index SII is defined as $SII = \sum_{i=1}^{3} l_i / \sum_{i=1}^{6} l_i$. In this way, we can account for differences in general interest between the children. Furthermore, a relative measure is particularly relevant for understanding educational choices, as it considers STEM interest relative to other fields. We call the measure *SII (relative) STEM interest*.

Our second measure for STEM interest is a book choice. Specifically, we offered children different books to choose from at the very end of the ex-post measurement. Two of the books were STEM-related ("The Amazing World of Technology," "The Earth") and the other two were not STEM-related ("Dinosaurs," "To Argue and to Be Friends"). Each child could choose one book as a present to be handed out at the end of the data collection. This measure is denoted *STEM book* and takes the value of 1 if a child chooses one of the STEM books and 0 otherwise.

Validation of our measure of STEM interest

To assess if relative STEM interest (SII) predicts educational choices, we validated it by correlating it with real-life choices, a common practice in economics (e.g., Dohmen et al., 2011; Sutter et al., 2013; Enke et al., 2022; Falk et al., 2023). 345 Austrian high-school graduates participated in an online survey measuring (1) *STEM interest*, as elicited in the main study described above, and (2) self-reported educational choices for tertiary education or training (either a study program or vocational training) at the individual level. We classify education choices (*STEM occupation*) using the classification system "ISCO-08" by the International Labor Organization.[10] In the next step, we use the ISCO-08 coding to classify occupational choices into STEM and non-STEM professions based on a study by the European Commission (2015). Following this classification, the STEM fields include life sciences, physics, mathematics and statistics, computing, engineering plus manufacturing, and processing. *STEM occupation* takes value one when a respondent's occupational choice falls into one of the respective STEM categories and zero otherwise. Participants for our validation study were high-school graduates from

[10] https://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

different regions across the country. We recruited via the schools' principals. The survey was conducted online using Qualtrics in spring 2021. More women (73%) than men (25%) or people who identify with neither gender (2%) responded.

In line with our findings in the sample of primary school children (see Section 4), we observe significant gender differences in relative STEM interest in the sample of high-school graduates. *STEM interest* (all (N=345): mean=0.423, sd=0.149; women (N=252): mean=0.385, sd=0.132; men (N=87): mean=0.535, sd=0.142) is significantly lower for women than for men (Mann-Whitney test, p<0.001).[11] Accordingly, fewer women than men want to specialize in STEM ($\chi^2(1)$=13.003, p<0.001), indicated by the variable *STEM occupation* (all (N=345): mean=0.290, sd=0.454; women (N=252): mean=0.240, sd= 0.428; men (N=87): mean=0.437, sd=0.499). Our measures of *STEM interest* and *STEM occupation* are highly significantly correlated with a correlation coefficient of 0.456 (p<0.001).[12] The correlation has imperfect explanatory power which is likely to be, first, due to the different scales of our survey instrument and the actual choices, and, second, measurement error leading to attenuation bias when predicting choices. However, the correlation of 0.456 is similar to correlations of survey instruments to predict experimental choices for economic preferences which range from 0.4 to 0.68 (Enke et al., 2022; Falk et al., 2023). Moreover, the gender differences in our STEM interest measure are consistent with our study results from elementary school children. This makes us confident that our measure of *STEM interest* is a well-suited proxy for STEM interest and presumably a valid predictor of occupational choices.

Elicitation of behavioral mechanism variables

Stereotypical thinking: We cover conscious (explicit) and subconscious (implicit) stereotypical thinking. To measure explicit stereotypes, we use a set of six questions such as "Who is more talented in math?" Children can answer on a five-point scale from "Girls are more talented" (value of 0) over "Both alike" (0.5) to "Boys are more talented" (1). We use the average across all answers as our measure of explicit stereotypes. To measure implicit gender stereotypes, we use the Implicit Association Test (IAT) (Greenwald et al., 1998), developed in social psychology for adults. We modify the IAT following Cvencek et al. (2011) to measure implicit gender stereotypes in kids.[13] In the test, children have to group words on either the right side to one

[11] A histogram of this variable by gender can be found in Figure A.2 in the Appendix.

[12] Box plots illustrating the association between the two measures can be found in Figure A.3 in the Appendix. Results from non-parametric tests (the Point-Biserial Correlation Coefficient is 0.456, with a confidence interval between 0.371 and 0.529) and logistic regressions (see Table A.1 in Appendix A) confirm the rather strong association between the two measures.

[13] The IAT is an established measure in psychology with a high internal consistency (Cronbach's alphas for the gender activities measure were α = .84, and α= .90, respectively (Cvencek et al., 2011)), but it has also been subject to criticism for its test-retest reliability. For instance, Rothermund and Wentura (2004) demonstrate that the choice of the juxtaposed target categories such as boys and girls in our study, can significantly affect the measured associations. Nevertheless, if we assume that gender is the crucial characteristic that influences STEM interest, then the measure remains relevant for our study. A comprehensive summary of the concerns related to the IAT can be found in Hall et al. (2015).

category (e.g., "male") or the left side to another category (e.g., "female") of the tablet screen. The first set of words included German names of females and males, and the second set included math-related and German-related items (e.g., "plus and minus," "textbook"). One word appears at a time on the screen and needs to be grouped correctly into categories, i.e., either "male" or "female" in the first set and "German" or "Math" in the second set. In the third and fourth sets, the categories are combined. In the third set, the combined categories conform to potential gender stereotypes (German/female, Math/male). In the fourth set, the categories are counter-stereotypical (German/male, Math/female). Children are asked to answer as quickly as possible in all four sets. To prevent them from answering randomly, they receive error messages on the screen in the first two trial sets if they are categorized incorrectly. To calculate the IAT score, we add up the response times in seconds in set four and subtract the sum of the response times in set three.

Growth mindset: The measure of a child's growth mindset is based on Blackwell et al. (2007). Instructors read out a scenario in which a child writes an exam in a new subject in school and gets a bad grade. Four different statements provide potential reasons for the bad grade. Two of the reasons are related to a lack of innate intelligence (fixed mindset) and two to the missing dedicated effort (growth mindset). Children have to rate the provided explanations on a Likert scale from one, "I do not agree at all", to five, "I totally agree." We created an index of the four answers ranging from zero to one, with a higher index value indicating a more pronounced growth mindset.

Competitiveness: We measure competitiveness with a modified version of the method by Niederle and Vesterlund (2011), similar to Sutter and Glätzle-Rützler (2015). We only implement two rounds due to time constraints. In the first round, children work on a math task under a piece-rate payment scheme for one minute. Each math task consists of four numbers that have to be added up. Children have to choose between three possible answers in a multiple-choice style. For each correct answer, they receive one point. Before we start the second round, children can decide whether they want to take part in a competition or whether they would prefer to be paid an individual piece rate for one point for each correct answer. When they choose the competition, they are anonymously matched with another child in the room. When children in the competition answer more questions correctly than the randomly-drawn competitor (i.e., they win), they receive two points per correct answer, and 0.5 points per correct answer otherwise (they lose).

(Relative) STEM confidence: To measure STEM confidence, we use the same job descriptions as for STEM interest, but here we ask children whether they would be confident working in these areas. We formulate the question in a way that can be easily understood by very young

children.[14] Children can answer on a 5-point Likert scale from "No, not at all" to "Yes, very much." For the analysis, we use a relative measure of STEM confidence. More precisely, we sum up the answers for the three STEM jobs and divide them by the sum of all answers. This results in the index *Stem confidence*, ranging from zero to one. Specifically, a value close to zero can be interpreted as very little relative STEM confidence, and a value of one as very high relative STEM confidence.[15]

3.4 Sample, procedures, and evaluation strategy

Procedures and data collection

To test our hypotheses, we collected individual-level data using survey measures and incentivized economic decisions. We collect data before ("baseline measurement") and after ("end measurement") a four-week intervention period. The baseline measurement enables us to check for (unlikely) a priori differences in relevant variables between the control and treatment groups and to test the effect of our behavioral mechanism variables on interest in STEM in general. We did not collect data on any variables related to professions or even STEM in the baseline measurements to avoid experimenter demand effects. The IAT could not be administered twice due to time constraints.

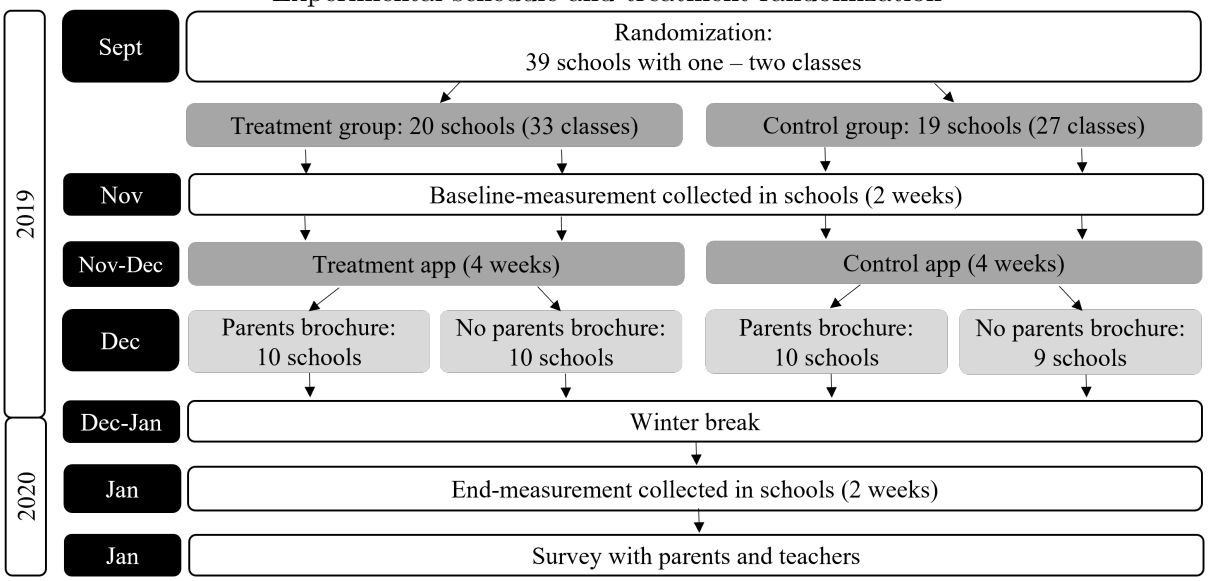
The baseline and end measurements were programmed with the software "oTree" (Chen et al., 2016). Children participated in the experiment with computer tablets. Sessions within class lasted approximately 60 minutes, and children were paid in the form of gifts, depending on performance (see below), with an average value of €1.50.

In Figure 2, we visualize the time frame of the experiment and its phases. First, we invited school principals and teachers (the randomization process is described below). After making appointments with each class, we went for the baseline data collection in November 2019. For all 39 schools, initial data collection took two weeks. The intervention period started on Monday after the first data collection, i.e., we had one group of schools starting one week earlier ("first cohort") than the other group of schools ("second cohort"). The intervention period for the second group ended just before the Christmas break. On the last day of the intervention, teachers handed out the brochures to the parents. In January, we came back to the schools for the final data collection. This was about four weeks after the end of the intervention period. A cooling-off period of several weeks renders the potential positive effects of the intervention stronger. Any significant effects would thus be an indication of a more lasting impact of the

[14] In German the question was "Wie sehr traust du dir das zu?" ("How much do you trust yourself in doing that?"). More generally, We tested decisions and questions in pilot sessions with teachers and children. If children showed difficulties in answering, we discussed alternative formulations, tested them, and changed them accordingly.

[15] We focus on STEM confidence in the following because we think that it is the more relevant variable in potentially determining STEM interest than math confidence that we also elicited.

FIGURE 2
Experimental schedule and treatment randomization



intervention.

For the baseline and end data collection, we sent two research assistants with a mobile lab to classes. We used dividers between the children to guarantee private decisions (see Appendix A, Figure A.1 for a picture of a session). One of the research assistants read the instructions for each task out loud. For some tasks, there were control questions on the tablet before the actual decisions to ensure that the children understood what they needed to do. In the baseline and end data collections, we use survey and lab-in-the-field experimental methods alike to elicit the mechanism variables. We informed the children at the very beginning of the session that they could earn points in several tasks/games. In the end, they would receive gifts depending on the relative number of points they collected on the experimental tasks within a class. The upper third of the class received bags with three gifts, the middle third two gifts, and the lower third, one gift. The bags contained small toys, stickers, and stationery such as pens. We showed a selection of gifts to the children in advance to make the scope of the incentives transparent. For our purpose, the coarse incentive scheme is sufficient and allows for easy implementation, given the restricted time of around one hour per class that was available for variable measurement ex-ante and ex-post of the intervention.

The intervention, i.e., the app, was introduced to the children at the end of the baseline data collection in class. When we were present at a treatment school, we briefly described the treatment app "Robitopia," and, similarly, when at a control school, the control app "Anton." To describe the apps, the same wording was used to make sure that children were equally excited to use the app in both conditions. After the introduction of the app, we handed out gift bags (depending on the performance in the baseline data collection) and letters to the parents. In

the letters, we explained the apps, how often they should be used, and how the children could log in using an individual code. The individual code was used to match the log-in data from the apps with the individual-level data from the baseline and end measurements in schools while preserving anonymity. To increase the take-up rate of the intervention as well as the usage time, we initiated a competition between classes by offering a price of €100 to the three classes with the most activity time on the apps. We sent weekly reminders to the teachers through SMS, prompting them to encourage children to utilize the app, while also reminding them about the prize.

Recruitment and randomization

We aimed at recruiting 40 elementary schools following the pre-registry, assigned randomly to the treatment and the control app. The sample size was determined by the resources (e.g., time, budget) available for the study.[16] We examined treatment effects in diverse geographical areas to derive tailored policy implications for potential scaling. We recruited schools in Vienna and Linz, urban areas, and in the rural region of "Muehlviertel," which is north of Linz. To select schools, we used a list of all the public schools in Vienna, Linz, and the Muehlviertel, provided by the Federal Ministry of Education, Science and Research. The public education administration offices supported the recruitment by providing a recommendation letter in Vienna and by making phone calls. However, participation was voluntary for the schools, i.e., we could not and did not want to force schools to take part in our study.

We formed school pairs based on geographical proximity and randomly assigned one school in each pair to the treatment.[17] Since parents cannot choose elementary schools for their children but are assigned to a school in their district/region by the school authority, we can assume that children within school pairs are more similar concerning their socioeconomic background than completely random school pairs.

Due to resource constraints, we could deal with a maximum of two classes within each school. When a recruited school had more than two classes in grade three, we chose the two classes with the lowest classroom numbers; that is, the number that identifies the room that a specific class in elementary school uses for all or the majority of lessons. In the following, we describe the recruitment for Vienna and Linz/Muehlviertel in more detail.

In Vienna, we focused the recruitment on the inner city districts (districts two to nine) for logistical reasons. Since there is a small possibility of systematic socioeconomic differences between children from these different districts, we stratified recruitment by district, i.e., the share of schools per district in our sample resembles the true share of schools in the district. More specifically, we randomly chose one or two schools in each district, depending on their size.

[16] With the given sample, we expected to be able to detect and effect size of about 0.24 standard deviations with a power of 80% and an alpha of 0.05, not considering improved precision when including control variables.

[17] In Vienna, we realized after recruitment that one school had already used the control app. This school was taken out of the randomization process and directly assigned to the treatment group. As a consequence, the other school in the pair was assigned to the control group.

In total, ten out of 61 schools were chosen. For these schools, we selected the nearest elementary school. If either the selected school or the nearest other school did not want to participate, we went down the list and contacted the second nearest neighboring school. We contacted a total of 35 schools of which 19 took part in our study.[18]

In the city of Linz and the rural area Muehlviertel, both in the state of Upper Austria, we sorted the school list by postal code and selected ten schools randomly from a total of 131 schools. For the ten schools, we consecutively contacted the schools geographically closest to the initial draw, similar to the recruitment in Vienna. If a school did not want to participate, we went down the list and contacted the next school on the list. We ended up contacting a total of 28 schools out of which 20 schools, with sufficient class size, agreed to take part in the study.[19] Four schools in Upper Austria are situated in the city of Linz and 16 in the region of Muehlviertel.

The parents' information brochure was handed out to half of the schools. More precisely, half of the control group and half of the treatment group were randomly assigned to the brochure condition.

3.5 Data

In total, 1,133 children participated in our RCT. We had to drop a few observations in our main analysis for the following reasons. Some children did not participate in both data collections, i.e., in the baseline and end measurements, mainly due to sickness absence. This reduces the sample by 155 observations. Moreover, we had to conduct a few sessions using pen and paper due to technical issues. This produced missing data for four children that did not answer all the survey questions. Seven children did not answer all socioeconomic questions, and three children had to leave sessions early. We excluded another two observations from children that stated that they had no interest in the six job descriptions at all. This leaves us with 962 observations for the main data analysis. The excluded children do not differ significantly from those included in the analysis in the variables that we specified in the pre-analysis plan (e.g., gender). Results are presented in Appendix B. We also replicate our main results (N=962) with the entire sample (N=1133) in the Online Appendix.

Table 1 presents a balance table, showing results from ordinary least squares regressions of the treatment dummy on the variables displayed in the first column, clustering standard errors

[18] Out of the 16 schools that eventually decided not to take part in the study, one school was excluded because it told us at the recruitment that classes had already used the control app and that it has concerns with the control app. The majority of the remaining schools did not give us a reason for why they did not want to participate. Those schools that provided reasons for their decline, stated a lack of time, concerns about the capability of some scholars to participate because of language barriers, and bad experiences with previous research studies (e.g., not getting informed about results after completion of the study).

[19] Out of the eight schools that did not participate, we excluded three schools that had fewer than 15 pupils in third grade (this was our ex-ante self-chosen cut-off number for classes to have enough statistical power in the analysis and to manage resources efficiently), three schools did not mention any reasons for not wanting to take part, and in one school the parents decided against participation.

by schools. The columns labeled "control" present the means of the control group. The columns labeled "difference" display the differences between the control group and the treatment group. The results show that the majority of the comparisons in observable variables between treatment and control are not significant at conventional levels with one exception. Girls in the treatment exhibit marginally significantly stronger explicit stereotypes than girls in the control group. In addition to the comparisons shown in Table 1, we also check for differences between the treatment and the control groups at the class level. Within classes, the share of girls is, on average, 50% and similar in the treatment and the control groups ($p = 0.852$, test of proportions).

The take-up rate of the treatment and the control app is comparable, with a slightly higher take-up rate in the treatment group. About 59% of the children in the control and 66% of the children in the treatment group have logged into the app at least once. However, this difference is not significant across treatment and control ($p = 0.233$, Wald test). Overall, the average exposure to the app measured in minutes played is similar (difference: -6.49 minutes, $p = 0.454$, Wald-test, see variable *Usage time*).[20] However, only looking at children who have used the app, the average usage time for the treatment app is with 172 minutes slightly lower than the suggested 200 minutes but equivalent to a daily usage time of 8.6 minutes. Average usage time in the treatment app is slightly lower than in the control app where it is 190 minutes, for those who have used the app at all ($p = 0.008$, Wald test).

Moreover, teachers reported the following specializations of their classes: 1) active learning and a focus on English, 2) multi-grade classroom, 3) multi-language classroom, 4) music, and 5) Montessori. The majority of the participating classes (74% in both treatment and control) did not have a specialization. The specializations are not distributed equally over treatment and control with specializations 1, 2, and 4 being over-represented in the treatment group, and 3) and 5) being over-represented in the control group. We, therefore, control for class specializations in the regression analyses.

4 Empirical results

We first test for gender differences in STEM interest and baseline gender differences in our behavioral mechanism variables (Hypothesis 3). Second, we present the main treatment effects, i.e., the effects of the intervention on STEM interest. Third, to better understand the behavioral mechanisms through which the treatment affected STEM interest, we examine how the mechanism variables are associated with interest in STEM, test whether the treatment alters the behavioral mechanism variables, and run a mediation analysis. In all regressions, we either use no control variables or the full set of control variables that include children's age, spoken

[20] While we cannot directly observe whether the children have been actively using the app while logged on, we have additional information on progress in terms of stars earned for finishing tasks and number of correct answers provided in the quizzes. Both of these variables are highly positively correlated with usage time (stars earned: ρ=0.84, p<0.001, correct answers: ρ=0.84, p<0.001).

TABLE 1
INTERNAL VALIDITY BALANCING TESTS

	FULL SAMPLE		GIRLS		BOYS	
	(1)	(2)	(3)	(4)	(5)	(6)
Variable	control	difference	control	difference	control	difference
Age[a]	8.49	0.07	8.45	0.09	8.53	0.04
	(0.69)	(0.06)	(0.62)	(0.07)	(0.75)	(0.09)
Language[b]	0.44	0.04	0.43	0.09	0.45	0.00
	(0.50)	(0.09)	(0.50)	(0.10)	(0.50)	(0.09)
Socio-economic status[c]	2.18	0.09	2.17	0.06	2.20	0.12
	(1.30)	(0.18)	(1.20)	(0.18)	(1.39)	(0.23)
Parents with STEM job[d]	0.29	0.01	0.34	-0.02	0.25	0.04
	(0.46)	(0.04)	(0.47)	(0.06)	(0.43)	(0.06)
Explicit stereotypes [e]	0.54	0.00	0.53	0.01*	0.55	-0.01
	(0.09)	(0.01)	(0.09)	(0.01)	(0.09)	(0.01)
Growth mindset[f]	0.57	-0.00	0.59	-0.01	0.56	0.00
	(0.15)	(0.01)	(0.14)	(0.01)	(0.16)	(0.02)
Competitiveness[g]	0.58	-0.02	0.47	0.00	0.68	-0.05
	(0.49)	(0.04)	(0.50)	(0.06)	(0.47)	(0.05)
Math confidence[h]	6.25	-0.08	5.18	-0.13	7.32	-0.06
	(7.58)	(1.09)	(7.09)	(1.30)	(7.92)	(1.09)
Take-up rate[i]	0.59	0.07	0.62	0.05	0.56	0.08
	(0.49)	(0.05)	(0.49)	(0.07)	(0.50)	(0.06)
Usage time[j]	112.53	-6.49	126.38	-10.19	98.54	-2.51
	(162.61)	(16.10)	(169.50)	(20.11)	(154.50)	(17.48)
Observations	418	962	210	480	208	482
Class Level Variables						
Rural area[k]	0.37	-0.05				
	(0.49)	(0.16)				
Public School[l]	1.00	-0.07				
	(0.00)	(0.07)				
STEM focus[m]	0.27	0.00				
	(0.55)	(0.13)				
Prior app usage[n]	0.46	-0.01				
	(0.51)	(0.14)				
Observations	27	58				

NOTE.— Numbers in parentheses indicate standard errors clustered by schools.
Variable definitions (see questionnaires/instructions for additional details): [a] *Children's age* in years, [b] *Spoken language at home*, 0=parents speak only German at home, 1=parents speak at least one other language than German at home, [c] *Proxy for socioeconomic status*, children estimate how many books there are at their homes, 0=0-10 books, 1=11-25 books, 2=26-100 books, 3=101-200 books, 4=more than 200 books, [d] 1=if parents indicated that they work in a STEM profession; 0 otherwise. $N = 531$ as not all parents answered the survey. We use ISCO-08 coding to identify STEM and non-STEM professions based on a study by the European Commission (2015), [e] Aggregate measure from six different questions ranging from 0 to 1, where 0.5 means that the child does not exhibit stereotypical thinking, values below 0.5 indicate anti-stereotypical thinking, and values above 0.5 stereotypical thinking in line with existing stereotypes, [f] Aggregate measure by Blackwell et al. (2007) based on four survey questions ranging from 0 to 1; increasing values indicate a more pronounced growth mindset, [g] 1=preferences for a math competition, 0=preference for piece-rate in the math exercise rather than competition, [h] Difference between the number of correct answers and the child's guessed number of correct answers in a math task [i] 1=child logged into the app at least once, 0=child did not log into the app at all, [j] Total number of minutes the child used the app (count of minutes starting with the log-in and stopping with the logout). [k] *Location of the school (urban or rural)*, 1= rural area, 0=urban area, [l] 1=school is a public school; 0 otherwise, [m] 1=teacher states that they actively promote STEM-related school activities; 0 otherwise, [n] 1=teacher states that she/he has previously used a learning app; 0 otherwise.

language at home, location of the school (urban or rural), class specialization, and a proxy for socio-economic status.[21]

4.1 Baseline gender differences

Girls may, in general, have a lower interest in STEM than boys already at the age of our participants. To test this, we focus on differences in the interest in STEM in our control group. We use data from the control group, as STEM interest is only measured after the intervention to avoid experimenter demand effects. We find that, indeed, girls choose a STEM book less often than boys (share of books chosen which are STEM books; girls: 0.56 and boys: 0.70, $p = 0.003$; test of proportions), and they also exhibit a substantially lower relative STEM interest (girls: 0.42 and boys: 0.61, $p < 0.001$; Wilcoxon-Mann-Whitney test).

Previous literature predominantly finds that girls are less confident and less competitive than boys in stereotypical male tasks (e.g., Buser et al., 2014; Dahlbom et al., 2011; Dreber et al., 2014; Sutter and Glätzle-Rützler, 2015). In line with this research, we find that, while about 66% of boys are willing to enter a math competition, only 47% of girls do so ($p < 0.001$, test of proportions). As a proxy for baseline confidence, we asked children to guess the number of tasks they had solved correctly in the math assignment. Whereas the actual performance of boys and girls is very similar, with around seven correctly solved tasks, boys overestimated their performance by about seven and girls by about five tasks ($p < 0.001$; Wilcoxon-Mann-Whitney test). Hence, we confirm gender differences in confidence and competitiveness (see Hypothesis 3c and Hypothesis 3d), and both girls' and boys' expectations of performance are far from being well-calibrated. Besides, we also observe gender differences in stereotypical thinking and growth mindset. Boys have significantly more pronounced stereotypical beliefs than girls (girls: 0.53, boys: 0.55, $p = 0.011$; Wilcoxon-Mann-Whitney test) and girls have a more pronounced growth mindset than boys in our sample (girls: 0.58 and boys 0.56, $p = 0.029$; Wilcoxon-Mann-Whitney test).

As girls have a lower initial level of STEM interest than boys, there is more room for increasing STEM interest (Hypothesis 2). Concerning our behavioral mechanism variables, we find small absolute but significant differences in the levels for growth mindset and stereotypical beliefs between boys and girls, but substantial differences in confidence and the willingness to enter a competition. Therefore, there is more scope of improvement in confidence and competitiveness for girls than for boys, making them potential mechanisms through which the app increases STEM interest (Hypotheses 3 and 4) particularly for girls.

[21] We present a set of model specifications, using a specific set of control variables based on children or teacher surveys to keep the main paper short. Results remain similar with different sets of control variables in a step-wise approach (see the Online Appendix). We do not control for the brochure intervention in the main part of the paper for brevity but demonstrate that including a dummy for the brochure intervention and an interaction term between the interventions does not affect our results in Table A.4.

4.2 Estimation of treatment effects

We use two different outcome variables to investigate the treatment effects on STEM interest, *STEM interest* and *STEM book*. *STEM interest* is defined as the interest in three STEM jobs relative to the sum of stated interest, i.e., the relative interest in the three STEM jobs *and* the three non-STEM jobs (see also the definition of SSI in Section 3). STEM interest is 0 if a child indicated no interest in all three STEM jobs and can be at a maximum of 1 if children indicated interest only in the STEM jobs but no interest in the non-STEM jobs. Consequently, a value of 0.5 indicates equal interest in STEM and non-STEM jobs. Overall, the children were slightly more interested in STEM jobs. The average score for STEM interest is 0.52, which is, given our large sample size, significantly larger than 0.5 ($p = 0.002$, sign-test). The variable *STEM book* is a dichotomous variable that takes the value 1 when a child chooses a STEM book and 0 for a non-STEM book. Overall, 62% of the children chose a STEM book, which is again significantly above 50% ($p < 0.001$, test of proportions).

In Table 2, we analyze the treatment effect on both measures for interest in STEM.[22] The upper panel shows the treatment effect on *STEM interest*. We find that boys are, on average, more interested in STEM jobs than girls. The relative interest for STEM is approximately 0.16 points lower for girls than for boys (see variable *Girls* in columns (1) and (2)).

Girls' STEM interest increases significantly in the treatment, compared to the control group by +0.04 to +0.06 points, which is equivalent to 0.21 to 0.31 standard deviations (see variable *Treatment* in columns (3) and (4)). Moreover, the treatment effect is significantly larger for girls than for boys (Wald tests at the bottom of the upper panel), which is in line with Hypothesis 2.[23] As a result, the gender gap in STEM interest decreases by 20%. In Figure 3, we illustrate this result by plotting the cumulative distribution functions (cdf) of STEM interest for girls and boys. The cdf for the treatment group (dashed black line) is below the control's cdf (solid grey line) in the left panel. This difference in distributions is also statistically significant ($p = 0.048$, global test of equality of distributions by Goldman and Kaplan (2018)). We do not find significant treatment effects in the pooled sample, except for a marginally significant effect when including all control variables, nor for boys separately (see variable *Treatment* in columns (1), (2), (5), and (6)).

The lower panel of Table 2 presents results from a linear probability model with the decision

[22] In Table A.5 in the Appendix, we present regressions controlling for parents' socioeconomic status, proxied by parents' level of education and their household income. These measures are based on parents' surveys instead of using the proxy based on the children's survey. Participation in the survey was voluntary, and 58% of the parents returned the survey. The size of the treatment coefficients is not much smaller than in the main regressions but they are estimated less precisely and are not statistically significant due to the smaller sample size. Our proxy for socio-economic status in the main analysis, measured by asking the children how many books they have at home, is significantly correlated with the parents' socio-economic status from the survey, i.e., level of education ($\rho = 0.36$, p-value < 0.001, $N = 562$) and household income ($\rho = 0.28$, p-value < 0.001, $N = 509$).

[23] While we use *relative* STEM interest as the outcome variable in the main part of the paper, we also provide an analysis looking at average interest in STEM jobs without accounting for interest in non-STEM-related jobs in Appendix C. We find that the increase in relative STEM interest is driven by an increase in interest in STEM jobs, whereas interest in non-related STEM jobs is not affected by the treatment.

Table 2
Direct treatment effects on interest in STEM

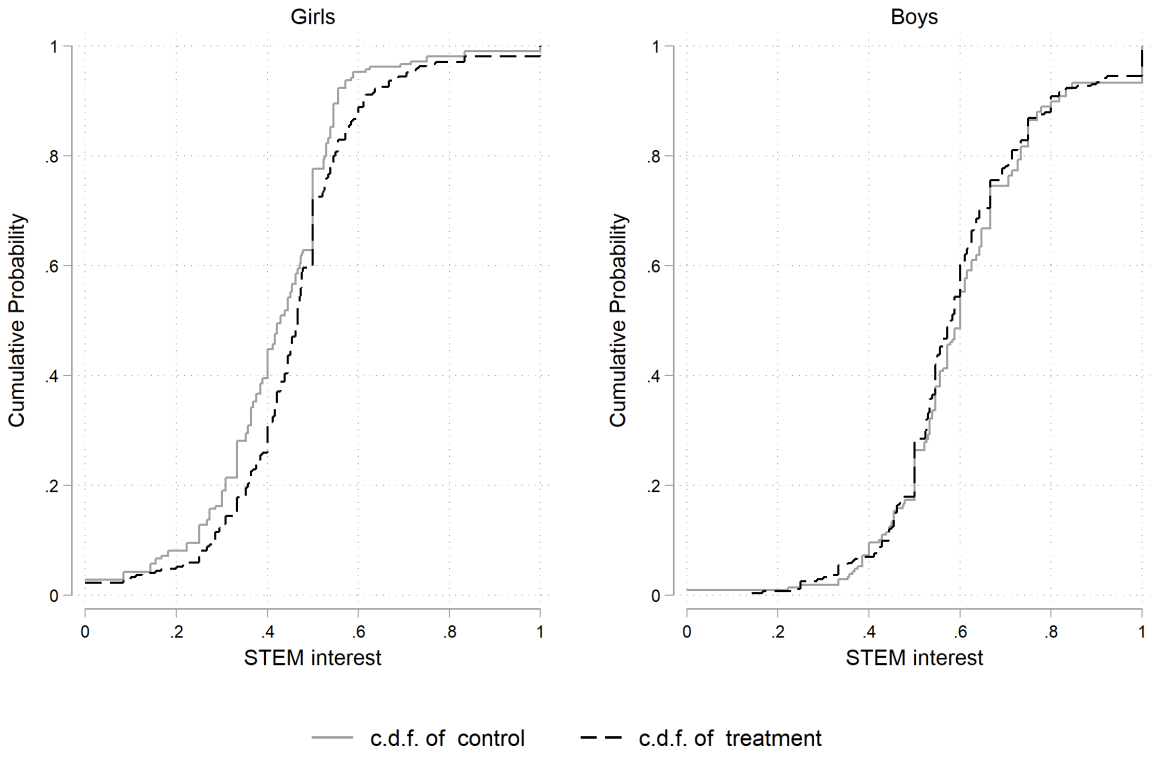
	STEM interest					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.016	0.026*	0.040**	0.060***	-0.009	-0.009
	(0.014)	(0.014)	(0.016)	(0.017)	(0.020)	(0.021)
	[0.500]	[0.168]	[0.049]	[0.009]	[0.649]	[0.706]
Girls	-0.163***	-0.163***				
	(0.012)	(0.013)				
Constant	0.594***	0.644***	0.417***	0.423***	0.608***	0.739***
	(0.012)	(0.061)	(0.010)	(0.101)	(0.013)	(0.062)
Observations	962	962	480	480	482	482
R-squared	0.205	0.221	0.016	0.044	0.001	0.048
Controls	no	yes	no	yes	no	yes
Gender differences in treatment effects		(3)-(5)	0.049**		(4)-(6)	0.055**
			(0.033)			(0.022)
	Chooses a STEM book					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	-0.029	0.001	0.020	0.031	-0.078	-0.045
	(0.043)	(0.044)	(0.054)	(0.061)	(0.059)	(0.056)
	[0.534]	[0.951]	[0.736]	[0.640]	[0.340]	[0.706]
Girls	-0.085**	-0.083**				
	(0.040)	(0.040)				
Constant	0.674***	0.804***	0.562***	0.746**	0.702***	0.802***
	(0.032)	(0.190)	(0.038)	(0.287)	(0.032)	(0.266)
Observations	962	962	480	480	482	482
R-squared	0.008	0.031	0.000	0.031	0.007	0.058
Controls	no	yes	no	yes	no	yes
Gender differences in treatment effects		(3)-(5)	0.097		(4)-(6)	0.104
			(0.073)			(0.073)

Note.— The upper panel shows OLS regressions with robust standard errors clustered at the school level and STEM interest as the dependent variable. The lower panel shows a linear probability model with robust standard errors clustered at the school level and the choice of a STEM book as the dependent variable. Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socio-economic status. Numbers in parentheses indicate standard errors. Numbers in brackets are Romano-Wolf p-values controlling for multiple hypotheses testing adjusted for the two different outcome variables. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

to choose a STEM book as the dependent variable.[24] We find that boys are generally more interested in STEM books than girls; they are 8 percentage points more likely to choose a STEM book (see variable *Girls* in columns (1) and (2)). We do not find significant treatment effects (see variable *Treatment*) for the book choice. We conclude that our treatment did not

[24]We present results from a logistic regression in Appendix A in Table A.3.

FIGURE 3
Cumulative distribution functions of STEM interest in control and in treatment



affect children's book choice. Thus, we focus on the relative *STEM interest* for the subsequent analyses on behavioral mechanisms. Since girls and boys respond differently to the treatment app, we continue with gender-disaggregated analyses. To sum up, we reject Hypothesis 1 and cannot reject Hypothesis 2 and summarize the first result as follows. We discuss the results in greater depth in Section 5.

Result 1:
Girls' are generally less interested in STEM professions than boys, and the treatment increases their STEM interest significantly. The treatment has no significant effect on children's STEM interest overall or for boys specifically.

We consider intention-to-treat effects in the main body of the paper, as the *offer* of the treatment is policy-relevant rather than the treatment effect on the treated. We discuss estimates for the treatment effect on the treated for the main results in section E of the Appendix. As expected, accounting for compliance, i.e., the take-up rate of 65%, the treatment effect size increases by 35%.

We also present heterogeneity analyses that controls for exposure levels, i.e., app usage time, in the Online Appendix. The treatment effect does not vary significantly with usage time. We

also focus on other potential heterogeneous effects following our pre-registration. The treatment seems to impact children with differing characteristics very similarly overall. Thus, we do not find heterogeneous treatment effects for children with or without migration background, captured by the spoken language at home, for low or high math ability, or for low or high tech affinity. Only school location has a differential treatment effect; we see a 5.4% higher increase in the treatment in rural than in urban areas. However, the effect vanishes as soon as we control for covariates. Details on the heterogeneity analyses can be found in Chapter 2 in the Online Appendix.

4.3 Behavioral mechanisms

4.3.1 Effect of behavioral mechanisms on STEM interest

To explore the effect of the behavioral mechanisms on interest in STEM, we conduct a correlation analysis between STEM interest and the mechanism variables presented in Table 3. For this analysis, we focus on the behavioral mechanisms that we measured in the baseline data collection, this is explicit stereotypical thinking, growth mindset, and competitiveness.

Particularly girls (but not boys) may have trouble expressing interest in STEM because of existing stereotypical beliefs and the resulting lack of identification (Kahn and Ginther, 2018; Miller et al., 2018). Consequently, we focus on girls in the regression that analyzes the effect of stereotypical thinking on STEM interest (column (1)), besides the regressions on the pooled data of the other behavioral mechanism variables in columns (2) and (3). We do not find support for Hypothesis 3a, that is, there is no evidence in our data that stereotypical thinking is associated with children's interest in STEM. Similarly, a growth mindset is not correlated with children's interest in STEM, rejecting Hypothesis 3b.

We find that competitiveness is significantly positively correlated with STEM interest. The index for relative STEM interest is 0.03 points or 0.16 standard deviations higher for children who decide to enter a competition ($p = 0.029$, Wald test, Table 3). This finding corroborates results from recent studies that demonstrate the importance of competitive aptitudes for education choices toward STEM fields (Buser et al., 2017, 2021). In Appendix A, we disaggregate the underlying model specifications from Table 3 by gender (see Table A.6). We find that the general positive effect of competitiveness in the regression is driven by girls. For girls that choose to compete, STEM interest increases by 0.04 points ($p = 0.019$, Wald test), and we do not find any effects from competitive aptitude for boys.

Since confidence is only assessed post-intervention, baseline measures cannot be used for measuring the impact for STEM interest. However, analyzing the control group reveals a strong correlation between confidence and interest (ρ=0.85, p<0.001, N=418).

Result 2:

Children's competitiveness and confidence are positively associated with children's interest in STEM. We do not find evidence that stereotypical thinking and a growth mindset are associated

Table 3
Determinants of STEM interest (baseline mechanism variables)

	(1) Girls	(2) Pooled	(3) Pooled
Girls		-0.164***	-0.159***
		(0.013)	(0.013)
Explicit stereotypes	0.035		
	(0.073)		
Growth mindset		0.018	
		(0.038)	
Competitiveness			0.025**
			(0.011)
Constant	0.414***	0.642***	0.643***
	(0.108)	(0.064)	(0.056)
Observations	480	962	962
R-squared	0.019	0.218	0.222
Controls	yes	yes	yes

Note.— OLS regressions with robust standard errors clustered at the school level. STEM interest is the dependent variable. Behavioral mechanism variables are measured at the baseline data collection. Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school, class specialization, and a proxy for socio-economic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

with STEM interest.

4.3.2 Treatment effect on behavioral mechanism variables

In Table 4, we examine whether the treatment was successful in altering the behavioral mechanism variables for boys and girls. We provide an analysis for the pooled sample in Appendix A (Table A.2). In the regression analysis, we incorporate controls for baseline levels of the behavioral mechanism variables, where available, in addition to the general controls.

In columns (1)-(2) and (6)-(7), we investigate the treatment effects on stereotypical thinking and differentiate between explicit stereotypical thinking and implicit stereotypical thinking (IAT). We find no significant effect neither on explicit or implicit stereotypes when looking at boys and girls separately. However, there is a significant decrease in explicit stereotypes in the pooled sample (see Table A.2).

In columns (3) and (8), we test whether the treatment fosters children's growth mindset. We find that the treatment significantly increases the levels for girls' growth mindset, but there is no effect on boys. More precisely, the measure for girls' growth mindset increases by 0.026 points through the treatment, equivalent to an increase of 0.17 standard deviations ($p = 0.021$, Wald test).

Columns (4) and (9) present linear probability models on the decision to enter the math competition (*Compet.*). We find that the chance of a girl entering the competition is 7.4 percentage points higher in the treatment than in the control group ($p = 0.056$, Wald test). We do

not find a significant effect on boys' willingness to enter the math competition, and the relative increase is significantly larger for girls than for boys (see the bottom row in Table 4). The large treatment effect on girls is also reflected in a marginally significant increase in competitiveness in the pooled sample (see Table A.2). In the pooled sample, we find that children in the treatment group are 3 percentage points more likely to choose the competition (and not piece-rate) than children in the control group ($p = 0.094$, Wald test).

Finally, columns (5) and (10) investigate the treatment effects on children's relative STEM confidence. Girls' STEM confidence increases significantly in the treatment group (difference: 0.047 points or 0.28 standard deviations, $p = 0.002$, Wald test), and the effect for girls is significantly larger than the effect for boys. Interestingly, we find a marginally significant negative effect on boys' STEM confidence (difference: -0.033 points or 0.19 standard deviations, $p = 0.068$, Wald test). Note that STEM confidence is a relative measure of confidence in STEM fields to confidence in non-STEM areas (see Section 3). The negative effect may, therefore, originate from the treatment building up boys' confidence in social jobs rather than losing confidence in STEM. The data provides suggestive evidence in line with this argument but the effect is not significant: In the treatment group, boys' confidence in non-STEM subjects and social jobs increases by 0.13 points on a Likert scale from 0–4 ($p = 0.133$, Wilcoxon-Mann-Whitney test). In contrast, boys' confidence in STEM subjects was not significantly affected by the treatment (-0.08 points, $p = 0.423$, Wilcoxon-Mann-Whitney test).

To sum up, the intervention decreased children's explicit stereotypes. Focusing on girls, we find that their growth mindset, their willingness to enter a competition, and their STEM confidence increase significantly. We conclude that the treatment was particularly successful in affecting behavioral mechanism variables for girls in the expected direction.

TABLE 4
TREATMENT EFFECTS ON BEHAVIORAL MECHANISM VARIABLES BY GENDER

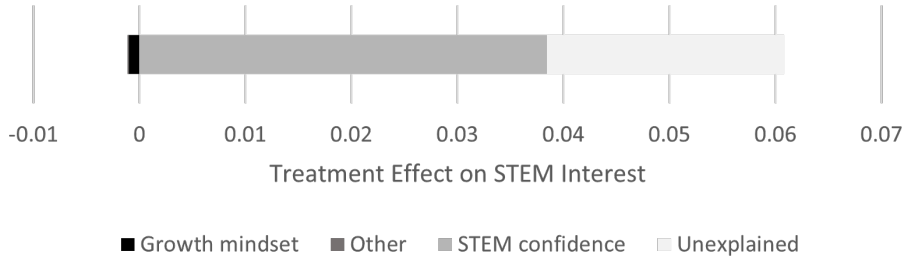
	GIRLS					BOYS				
	(1) Explicit stereotypes	(2) IAT	(3) Growth Mindset	(4) Compet.	(5) STEM Confidence	(6) Explicit stereotypes	(7) IAT	(8) Growth Mindset	(9) Compet.	(10) STEM Confidence
Treatment	-0.009	0.292	0.026**	0.072*	0.047***	-0.015	-0.248	0.008	-0.015	-0.033*
	(0.006)	(0.817)	(0.011)	(0.036)	(0.014)	(0.010)	(0.747)	(0.015)	(0.032)	(0.017)
	[0.356]	[0.802]	[0.149]	[0.337]	[0.040]	[0.505]	[0.852]	[0.861]	[0.911]	[0.366]
Math performance				0.012					0.008	
				(0.007)					(0.006)	
Baseline levels										
Explicit stereotypes	0.035					0.094**				
	(0.046)					(0.046)				
Growth mindset			0.260***					0.294***		
			(0.039)					(0.048)		
Competitiveness				0.371***					0.418***	
				(0.048)					(0.039)	
Constant	0.405***	1.970	0.375***	-0.441*	0.463***	0.534***	5.223*	0.401***	-0.097	0.724***
	(0.049)	(2.067)	(0.091)	(0.248)	(0.078)	(0.047)	(3.070)	(0.095)	(0.303)	(0.093)
Observations	480	480	480	480	480	482	482	482	482	482
R-squared	0.056	0.023	0.102	0.225	0.040	0.033	0.011	0.121	0.177	0.048
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Gender differences	0.009	0.525	0.022	0.091*	0.063***					
in treatment effects	(0.011)	(0.583)	(0.016)	(0.054)	(0.021)					

NOTE.— OLS regressions with behavioral mechanisms variables as the dependent variables. Columns (4) and (10) show linear probability models. Columns (1) to (5) only include girls. Columns (6) to (10) only include boys. Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socio-economic status. Numbers in parentheses indicate standard errors clustered by schools. Numbers in brackets are Romano-Wolf p-values controlling for multiple hypotheses testing. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

4.3.3 Mediation analysis

To learn more about the behavioral mechanisms and their explanatory power for the treatment effect (Hypothesis 4), we conduct a mediation analysis. For this, we decompose the average treatment effect in a direct effect of the treatment on STEM interest and indirect effects through changes in the behavioral mechanism variables. We focus on girls as this is the group that was affected by the treatment.[25]

FIGURE 4
Total treatment effect on STEM interest (Girls)



NOTE.— Decomposition of the treatment effect for girls ($N = 480$). The treatment effect is estimated using OLS regressions with standard errors clustered by schools. Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school, class specialization, and a proxy for socioeconomic status. The size of the bar corresponds to the unconditional treatment effect, and each section of the bar represents the share of the treatment effect that is explained by our behavioral mechanisms. All mechanisms are included separately but combined with "other" after the estimation, as competitiveness and stereotypes exhibit no mediating effect.

As a causal interpretation would require the assumption of sequential ignorability to hold, which cannot be tested, we do not claim to identify causal effects for this mediation analysis. However, the decomposition of the treatment effect still provides interesting insights on the relative importance of each of the mechanism variables (see for example Bandiera et al., 2020; Carneiro et al., 2020). We decompose the effect using the method by Gelbach (2016). This method treats each of the mechanisms as an omitted variable and uses the formula for omitted variable bias to decompose the total treatment effect into a share explained by the mechanisms (indirect) and an unexplained (direct) treatment effect. Thereby, it takes into account the fact that the mechanisms might be correlated with each other. Figure 4 shows the results of this decomposition analysis, focusing on girls. Similar to the results from the correlation, we find that the treatment effect is mainly driven by a change in STEM confidence. However, we do not find a significant effect of the willingness to enter a competitive environment, nor of less stereotypical thinking. The growth mindset's coefficient on STEM interest is negative but not statistically significant. Around 30% of the treatment effect remains unexplained by our behavioral mechanism variables.

[25] Before data collection in our preregistration, we hypothesized gender disparities in STEM interest, anticipating more pronounced treatment effects for girls than boys (see also Hypothesis 2). However, we did not anticipate that boys would remain unaffected by the treatment; rather, we expected them to enhance their already relatively high inherent STEM interest even further. Ultimately, we discovered that the intervention had no discernible impact on boys' STEM interests. Therefore, our preregistered mediation analysis, reliant on detecting treatment effects, are only employed for girls.

Result 3:
The increase in girls' STEM confidence explains a large part of the treatment effect. While the treatment app increases competitive aptitudes, we do not find a mediating effect from competitiveness on STEM interest.

5 Discussion

Assessing results

Outcomes: Although we detect a positive effect on the relative STEM interest measure, there is no evidence to suggest that the treatment prompted children to select a STEM book after using the treatment app. We can only speculate on the reasons for the absence of a treatment effect on the book choice. It is possible that children's book choice is not driven by their interest in a topic but by other criteria such as the attractiveness of the book cover. Especially, as children were facing a limited selection of books to choose from (we offered two STEM and two non-STEM books), other characteristics than the STEM focus could have driven the children's choice, muting a potential increase in interest for STEM books, on average. Another reason might be that, while our survey measure captures professional interests, such interest does not necessarily translate into a preference to read a STEM-related book. For example, children in the treatment group might have had the desire to read something different after exposure to STEM topics in the treatment. We conclude that the book choice might not be a good proxy for a general inclination towards STEM in our setting.

Mechanisms: We observe a positive correlation between girls' confidence levels and their interest in STEM, with confidence as the primary mediator of the treatment effect in our study. This finding is in line with previous studies showing that women demonstrate lower performance, contribute less to discussions, are less confident in their performance, and enter competitive environments less frequently in stereotypical male tasks (Günther et al., 2010; Niederle and Vesterlund, 2010; Balafoutas and Sutter, 2012; Coffman, 2014; Bordalo et al., 2019; Exley and Kessler, 2022; Gneezy et al., 2003; Saltiel, 2023), which can reduce their perseverance and interest in these tasks (Buser and Yuan, 2019). Building upon our findings regarding the significance of girls' confidence in nurturing their inclination towards STEM, the results imply a strategic emphasis on fostering non-economic prerequisites, such as STEM confidence. These factors are crucial for enabling women to cultivate STEM interests, persist, and excel within competitive STEM environments, as well as other competitive arenas. The lower self-confidence among girls compared to boys could potentially pose a double burden, as girls may require even higher levels of self-assurance to sustain their interest in STEM fields compared to boys. This additional self-confidence becomes imperative for girls to counteract stereotypical and societal expectations that may undermine their pursuit of STEM interests.

Competitiveness is positively associated with STEM interest in our study. The treatment increases competitiveness by seven percentage points. This effect size may not be pronounced enough to be a significant mediator. Competitiveness should not be dismissed as unimportant and be addressed more directly in the future when designing interventions to increase girls' interest in STEM.

Interestingly, we do not find a strong association between the other two mechanism variables of stereotypical thinking and children's growth mindset with interest in STEM. While previous literature focused on the influence of these mechanisms on career choices in older children and adults, there is little evidence on the effect of career aspirations at an early age, comparable to the age of our study participants (e.g., Gottfredson, 1981; Ambady et al., 2001). We might lack effects for the growth mindset because the belief that one's skills are malleable by working hard and the resulting increase in the willingness to seek challenging tasks might not be connected to STEM interest in young children, yet. Their perception of STEM fields and subjects as "hard and challenging" might also be less pronounced compared to adolescents (Archer et al., 2020). Thus, believing that one can actually learn from challenging tasks might be less important for young children's interest in STEM.

Obviously, our study does not cover the universe of potential behavioral mechanisms, particularly as it is the first to specifically investigate inclination towards STEM. Furthermore, it is natural that some of the variables are correlated. Future experimental studies could systematically manipulate various behavioral drivers within interventions to identify their individual impact more precisely. Moreover, future studies could delve deeper into the behavioral mechanisms of competitiveness and STEM confidence, possibly by varying intervention components such as feedback or rewards to determine their effectiveness.

Addressing potential identification threats

For our study design, we took precautions to mitigate identification threats including a potential demand effect. First, we use a control group with a similar app and introduce it in the exact same way as the treatment app, e.g., the oral description in class to the children and the information letters to teachers and parents about the app are identical for the treatment and the control group. The pairwise randomization further contributes to a clear identification. Second, we made sure not to refer to STEM in any way before and during the intervention phase. The apps were introduced as a way to "facilitate learning in an environment without social prejudices." Neither the term "STEM" nor anything related to the study's focus was mentioned in either the control or the treatment group until the end measurement. Third, we implemented several measures to prevent participants from being aware of the study's purpose until its conclusion. This included providing similar information to research assistants, principals, and other involved parties and disclosing the main variable of interest only at the study's debriefing after the end measurement. Hence, any social desirability bias in responses is very unlikely.

Timing of the intervention and persistence of outcomes

Previous research in economics has primarily focused on interventions aimed at adolescents shortly before they make decisions regarding tertiary education (e.g., Porter and Serra, 2020; Breda et al., 2023). However, our study highlights that STEM interest and underlying behavioral mechanisms contributing to the gender gap in STEM emerge early in life. These predispositions are likely to persist and influence later life outcomes (Francesconi and Heckman, 2016). Therefore, interventions targeting STEM interest and underlying reasons should commence earlier, as they often lay the groundwork for subsequent educational and occupational choices. Our approach exemplifies such an early intervention.

While the sustainability of the intervention's impact on STEM interest requires validation in long-term studies, our confidence in its effectiveness is bolstered by the observed effects persisting even after a significant interval between the intervention period and the final measurement. However, the long-term endurance of these changes remains a pertinent research question that warrants further investigation.

Although investigating the direct relationship between our treatment and investment in STEM education would have been intriguing, strict data protection regulations in Austria prevent us from linking individual-level experimental data to children's educational choices or outcomes. While we acquired aggregated administrative data on secondary school choices at the school level, discerning treatment effects from this data is challenging as only extremely large effects could be detected at the school level. Further details regarding the data and its analysis are available in the Online Appendix.

Our measure of STEM interest aligns with existing literature demonstrating the influential impact of educational interventions on subsequent choices. Economists have demonstrated leverage effects in various educational domains, such as negotiation training for girls leading to better-quality and longer-term education (Ashraf et al., 2020), grit training in schools enhancing math scores (Alan et al., 2019), and brief visits from female STEM role models influencing study program choices (Breda et al., 2023). Notably, these studies share intervention periods similar to or shorter than our own. Furthermore, our auxiliary validation study affirms the association between higher STEM interest and actual choices for STEM tertiary education, corroborating findings from another study (Drescher et al., 2020). This collective evidence underscores the validity of our STEM interest measure and its potential to translate into meaningful educational choices.

Assessment of expected effects when scaling up the intervention

The results of this study caught the attention of policymakers in Austria. They see the potential to implement our intervention on a larger scale by including it in the school syllabus. However, research suggests that there can be "scalability problems" as soon as one rolls out an intervention that has been tested on a significantly smaller scale (Al-Ubaydli et al., 2017).

interventions may fail to scale due to false-positive results in the initial study. Given our pre-registration, rigorous experimental design, and robustness checks, this scenario seems unlikely. Furthermore, our control group actively engaged with electronic devices and an online app, potentially leading to an underestimation of the total treatment impact, which encompasses exposure to STEM-related content and active electronic device interaction.

Second, the study sample may differ from the general targeted population. While our study drew a random sample from rural and urban regions in Austria and randomly assigned treatment and control groups within this sample, these regions may not fully represent the entire country. Nonetheless, we believe other regions may be comparable concerning culture and children's educational and social background.

Third, the context matters. In a study context, usually the principal investigators are in charge and invest time and effort to achieve high take-up rates of the intervention and comply with their investigation protocol (e.g., Grosch et al., 2023). However, when interventions are implemented on a larger scale, institutions typically assume responsibility for monitoring and managing the intervention, often overseeing numerous pupils. These institutions may face challenges in dedicating resources to closely monitor implementation, unlike researchers who can provide more intensive oversight. In our study, however, the principal investigators also implemented relatively loose monitoring, partly due to the large sample size and the resource-intensive nature of data collection. Moreover, our intervention was an out-of-school intervention in which children and their parents could choose whether they wanted to use the app or not. The research team had only two major points of contact with the children - one at the beginning and one at the end of the study for about an hour each. When, in the future, public authorities recommended the app or even incorporate it into the school curriculum, the effect could thus be larger than in our study. Moreover, the nature of the intervention allows for easy experimentation with the app's content, offering opportunities to improve outcomes for the target group. Therefore, we anticipate that our results would scale up effectively.

6 Conclusion

Increasing girls' inclination towards STEM fields from an early age could be pivotal in addressing the gender disparity in STEM careers and mitigating the gender income gap. Research findings indicate that gender disparities in ability or performance in STEM fields play a modest role in explaining gender differences in STEM educational and career trajectories (e.g., Hyde et al., 2008; Saltiel, 2023). Moreover, empirical evidence suggests that women's occupational sorting can be explained by academic skills and comparative advantages to a lesser extent than men's (Saltiel et al., 2019; Aucejo and James, 2021). Thus, focusing on the inclination towards STEM, rather than solely on STEM ability or performance, could be crucial in closing the gender gap in STEM.

This study provides encouraging evidence that interventions can impact individuals' incli-

nation towards STEM, representing, to the best of our knowledge, the first investigation of the malleability of STEM interests through a behavioral intervention. Our web application significantly boosted girls' interest in STEM and, thereby, reduced the gender STEM gap by 20%. We observe minimal variations in treatment effects across groups with different characteristics, indicating that our treatment has a similar impact on children, regardless of their ex ante technological affinity, migration background, or math ability. Moreover, we extend the literature on gender differences in STEM by examining behavioral mechanisms as potential underlying reasons for the gender STEM interest gap. We find that girls exhibit lower initial levels than boys for some of these behavioral mechanisms such as STEM confidence and competitiveness. In line with related research, we find that competitiveness is positively associated with STEM interest and our treatment successfully increases competitiveness. We see our results as a proof of concept: a digital intervention that addresses behavioral mechanisms can decrease the gender STEM interest gap. More research is needed on similar interventions to understand whether specific parameters or specific types, e.g., out-of-school or in-school interventions, work perhaps even better.

From a policy perspective, a digital intervention that increases interest in STEM is appealing due to its scalability, i.e., it can be distributed, extended, and tested with relative ease. Moreover, it can be more cost-effective and generally easier to integrate with school teaching, compared to more traditional interventions such as role models visiting classrooms or information events for parents. Consequently, we believe that similar interventions offer a promising approach to narrowing the gender gap in STEM and advancing gender equality in education and the job market.

References

Al-Ubaydli, O., J. A. List, and D. L. Suskind (2017). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review 107*(5), 282–286.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics 134*(3), 1121–1162.

Alan, S. and S. Ertac (2018a). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy 126*(5), 1865–1911.

Alan, S. and S. Ertac (2018b). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association 17*(4), 1147–1185.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Ambady, N., M. Shih, A. Kim, and T. L. Pittinsky (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science 12*(5), 385–390. PMID: 11554671.

Archer, L., J. Moote, E. Macleod, B. Francis, and J. DeWitt (2020). Aspires 2: Young people's science and career aspirations, age 10–19.

Ashraf, N., N. Bau, C. Low, and K. McGinn (2020). Negotiating a better future: How interpersonal skills facilitate intergenerational investment. *Quarterly Journal of Economics 135*(2), 1095–1151.

Aucejo, E. and J. James (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy 129*(10), 2905–2946.

Balafoutas, L. and M. Sutter (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science 335*(6068), 579–582.

Bandiera, O., N. Buehren, R. Burgess, M. Goldstein, S. Gulesci, I. Rasul, and M. Sulaiman (2020, January). Women's empowerment in action: Evidence from a randomized control trial in africa. *American Economic Journal: Applied Economics 12*(1), 210–259.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. NBER Working Paper 26993, National Bureau of Economic Research.

Berkowitz, T., M. W. Schaeffer, E. A. Maloney, L. Peterson, C. Gregor, S. C. Levine, and S. L. Beilock (2015). Math at home adds up to achievement in school. *Science 350*(6257), 196–198.

Bertrand, M., C. Goldin, and L. F. Katz (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics 2*(3), 228–255.

Bettinger, E., S. Ludvigsen, M. Rege, I. F. Solli, and D. Yeager (2018). Increasing perseverance in math: Evidence from a field experiment in norway. *Journal of Economic Behavior & Organization 146*, 1–15.

Bettinger, E. P. and B. T. Long (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review 95*(2), 152–157.

Bian, L., S.-J. Leslie, and A. Cimpian (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science 355*(6323), 389–391.

Blackwell, L. S., K. H. Trzesniewski, and C. S. Dweck (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development 78*(1), 246–263.

Blau, F. D., P. Brummund, and A. Y.-H. Liu (2013). Trends in occupational segregation by gender 1970-2009: Adjusting for the impact of changes in the occupational coding system. *Demography 50*(2), 471–494.

Blau, F. D. and L. M. Kahn (2000). Gender differences in pay. *Journal of Economic Perspectives 14*(4), 75–99.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–773.

Breda, T., J. Grenet, M. Monnet, and C. Van Effenterre (2023). How effective are female role models in steering girls towards stem? evidence from french high schools. *The Economic Journal*, uead019.

Brenøe, A. A. and U. Zölitz (2020). Exposure to more female peers widens the gender gap in stem participation. *Journal of Labor Economics 38*(4), 1009–1054.

Brocas, I. and J. D. Carrillo (2021). Steps of reasoning in children and adolescents. *Journal of Political Economy 129*(7), 000–000.

Brown, C. and M. Corcoran (1997). Sex-based differences in school content and the male-female wage gap. *Journal of Labor Economics 15*(3), 431–465.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics 129*(3), 1409–1447.

Buser, T., M. Niederle, and H. Oosterbeek (2021). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. NBER Working Paper 28916, National Bureau of Economic Research.

Buser, T., N. Peter, and S. C. Wolter (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. *American Economic Review 107*(5), 125–130.

Buser, T. and H. Yuan (2019). Do women give up competing more easily? Evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics 11*(3), 225–252.

Cappelen, A., J. List, A. Samek, and B. Tungodden (2020). The effect of early-childhood education on social preferences. *Journal of Political Economy 128*(7), 2739–2758.

Card, D. and A. A. Payne (2021). High school choices and the gender gap in stem. *Economic Inquiry 59*(1), 9–28.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *Quarterly Journal of Economics 134*(3), 1163–1224.

Carneiro, P., L. Kraftman, G. Mason, L. Moore, I. Rasul, and M. Scott (2020). The impacts of a multifaceted pre-natal intervention on human capital accumulation in early life. IZA Discussion Papers 13955, Institute of Labor Economics (IZA).

Cedefop (2015). Skills, qualifications and jobs in the eu: The making of a perfect match? Evidence from cedefop's european skills and jobs survey. Technical Report No. 103, Luxembourg: Publications Office. `https://www.cedefop.europa.eu/files/3072_en.pdf`.

Chen, D. L., M. Schonger, and C. Wickens (2016). oTree - An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics 129*(4), 1625–1660.

Cohodes, S. R., H. Ho, and S. C. Robles (2022). Stem summer programs for underrepresented youth increase stem degrees. NBER Working Paper 30227, National Bureau of Economic Research.

Cook, P. J., K. Dodge, G. Farkas, J. Fryer, Roland G, J. Guryan, J. Ludwig, S. Mayer, H. Pollack, and L. Steinberg (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in chicago. NBER Working Paper 19862, National Bureau of Economic Research.

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives 15*(2), 213–238.

Cvencek, D., A. N. Meltzoff, and A. G. Greenwald (2011). Math–gender stereotypes in elementary school children. *Child Development 82*(3), 766–779.

Dahlbom, L., A. Jakobsson, N. Jakobsson, and A. Kotsadam (2011). Gender and overconfidence: Are girls really overconfident? *Applied Economics Letters 18*(4), 325–327.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review 95*(2), 158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources 42*(3), 528–554.

Delaney, J. M. and P. J. Devereux (2019). Understanding gender differences in stem: Evidence from college applications. *Economics of Education Review 72*, 219–238.

Deming, D. J. and K. Noray (2020). Earnings dynamics, changing job skills, and stem careers. *Quarterly Journal of Economics 135*(4), 1965–2005.

DeWitt, J., J. Osborne, L. Archer, J. Dillon, B. Willis, and B. Wong (2013). Young children's aspirations in science: The unequivocal, the uncertain and the unthinkable. *International Journal of Science Education 35*(6), 1037–1063.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association 9*(3), 522–550.

Dreber, A., E. von Essen, and E. Ranehill (2014). Gender and competition in adolescence: Task matters. *Experimental Economics 17*(1), 154–172.

Drescher, K., S. Haeckl, and J. Schmieder (2020). STEM careers: Workshops using role models can reduce gender stereotypes. *DIW Weekly Report 10*(13), 163–172.

Eagly, A. H., W. Wood, and A. B. Diekman (2000). Social role theory of sex differences and similarities: A current appraisal. *Developmental Social Psychology of Gender 12*, 174.

Eble, A. and F. Hu (2020). Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review 77*, 101994.

Enke, B., R. Rodriguez-Padilla, and F. Zimmermann (2022). Moral universalism: Measurement and economic relevance. *Management Science 68*(5), 3590–3603.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature 58*(4), 897–996.

European Commission (2015). Does europe need more stem graduates. Technical Report No. 120/01, Luxembourg: Publications Office. `https://op.europa.eu/en/publication-detail/-/publication/60500ed6-cbd5-11e5-a4b5-01aa75ed71a1`.

European Institute for Gender Equality (2017). Gender segregation in education, training and the labour market. Technical Report No. 14624/17. `https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vkjm9hyf22z0`.

Eurostat (2018). Gender pay gap statistics. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Gender_pay_gap_statistics#Gender_pay_gap_levels_vary_significantly_across_EU`.

Eurostat (2019). Graduates in tertiary education, in science, math., computing, engineering, manufacturing, construction, by sex - per 1000 of population aged 20-29. `https://data.europa.eu/euodp/en/data/dataset/HlUaM7qrjUXzAAfE6UXbKg`.

Exley, C. L. and J. B. Kessler (2022). The gender gap in self-promotion. *Quarterly Journal of Economics 137*(3), 1345–1381.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Fehr, E., H. Bernhard, and B. Rockenbach (2008). Egalitarianism in young children. *Nature 454*(7208), 1079–1083.

Flory, J. A., A. Leibbrandt, and J. A. List (2015). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *Review of Economic Studies 82*(1), 122–155.

Francesconi, M. and J. J. Heckman (2016). Child development and parental investment: Introduction. *Economic Journal 126*(596), F1–F27.

Fryer Jr, R. G. and S. D. Levitt (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics 2*(2), 210–240.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics 34*(2), 509–543.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Ginther, D. K. and S. Kahn (2004). Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives 18*(3), 193–214.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics 118*(3), 1049–1074.

Goldman, M. and D. M. Kaplan (2018). Comparing distributions by multiple testing across quantiles or cdf values. *Journal of Econometrics 206*(1), 143–166.

Gottfredson, L. S. (1981). Circumscription and compromise: A developmental theory of occupational aspirations. *Journal of Counseling Psychology 28*(6), 545.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology 74*(6), 1464.

Grosch, K., S. Haeckl, H. Rau, and P. Preuss (2023). A guide to conducting school experiments: Expert insights and best practices for effective implementation. UiS Working Papers in Economics and Finance 2023/2, University of Stavanger.

Günther, C., N. A. Ekinci, C. Schwieren, and M. Strobel (2010). Women can't jump?—An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization 75*(3), 395–401.

Hall, W. J., M. V. Chapman, K. M. Lee, Y. M. Merino, T. W. Thomas, B. K. Payne, E. Eng, S. H. Day, and T. Coyne-Beasley (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American journal of public health 105*(12), e60–e76.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics 54*(1), 3 – 56.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science 312*(5782), 1900–1902.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010). The rate of return to the highscope perry preschool program. *Journal of Public Economics 94*(1), 114 – 128.

Heller, S. B. (2014). Summer jobs reduce violence among disadvantaged youth. *Science 346*(6214), 1219–1223.

Hermes, H., M. Huschens, F. Rothlauf, and D. Schunk (2021). Motivating low-achievers—Relative performance feedback in primary schools. *Journal of Economic Behavior and Organization 187*, 45–59.

Hyde, J. S., S. M. Lindberg, M. C. Linn, A. B. Ellis, and C. C. Williams (2008). Gender similarities characterize math performance. *Science 321*(5888), 494–495.

ILO (2019). A quantum leap for gender equality: for a better future of work for all. Technical report, International Labour Office.

Jiang, X. (2021). Women in stem: Ability, preference, and value. *Labour Economics 70*, 101991.

Joensen, J. S. and H. S. Nielsen (2016). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal 126*(593), 1129–1163.

Kahn, S. and D. Ginther (2018). Women and science, technology, engineering, and mathematics (stem). In *The Oxford Handbook of Women and the Economy*.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. NBER Working Paper 20749, National Bureau of Economic Research.

Kier, M. W., M. R. Blanchard, J. W. Osborne, and J. L. Albert (2014). The development of the stem career interest survey (stem-cis). *Research in Science Education 44*(3), 461–481.

Kollmayer, M., B. Schober, and C. Spiel (2018). Gender stereotypes in education: Development, consequences, and interventions. *European Journal of Developmental Psychology 15*(4), 361–377.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy 128*(2), 000–000.

Levanon, A., P. England, and P. Allison (2009, 12). Occupational feminization and pay: Assessing causal dynamics using 1950–2000 U.S. census data. *Social Forces 88*(2), 865–891.

List, J. A., R. Petrie, and A. Samek (2021). How experiments with children inform economics. NBER Working Paper 28825, National Bureau of Economic Research.

List, J. A., R. Petrie, and A. Samek (2023). How experiments with children inform economics. *Journal of Economic Literature 61*(2), 504–564.

List, J. A., A. Samek, and D. L. Suskind (2018). Combining behavioral economics and field experiments to reimagine early childhood education. *Behavioural Public Policy 2*(1), 1–21.

Mayo, M. J. (2009). Video games: A route to large-scale stem education? *Science 323*(5910), 79–82.

McNally, S. (2020). Gender differences in tertiary education: What explains stem participation? IZA Policy Paper 165, Institute of Labor Economics (IZA), Bonn.

Miller, D. I., K. M. Nolla, A. H. Eagly, and D. H. Uttal (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of us draw-a-scientist studies. *Child Development 89*(6), 1943–1955.

Murphy, R. and F. Weinhardt (2020). Top of the class: The importance of ordinal rank. *The Review of Economic Studies 87*(6), 2777–2826.

Niederle, M. and L. Vesterlund (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives 24*(2), 129–44.

Niederle, M. and L. Vesterlund (2011). Gender and competition. *Annual Review of Economics 3*(1), 601–630.

Nollenberger, N., N. Rodríguez-Planas, and A. Sevilla (2016). The math gender gap: The role of culture. *American Economic Review 106*(5), 257–261.

Olsson, M. and S. E. Martiny (2018). Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in Psychology 9*, 2264.

Oreopoulos, P., R. S. Brown, and A. M. Lavecchia (2017). Pathways to education: An integrated approach to helping at-risk high school students. *Journal of Political Economy 125*(4), 947–984.

Porter, C. and D. Serra (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics 12*(3), 226–254.

Rege, M., P. Hanselman, I. F. Solli, C. S. Dweck, S. Ludvigsen, E. Bettinger, R. Crosnoe, C. Muller, G. Walton, A. Duckworth, and D. S. Yeager (2021). How can we inspire nations of learners? An investigation of growth mindset and challenge-seeking in two countries. *American Psychologist 76*(5), 755.

Reuben, E., P. Sapienza, and L. Zingales (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences 111*, 4403–4408.

Rothermund, K. and D. Wentura (2004). Underlying processes in the implicit association test: dissociating salience from associations. *Journal of Experimental Psychology: General 133*(2), 139.

Saltiel, F. (2023). Multi-dimensional skills and gender differences in stem majors. *The Economic Journal 133*(651), 1217–1247.

Saltiel, F. et al. (2019). What's math got to do with it? Multidimensional ability and the gender gap in stem. In *2019 Meeting Papers*, Volume 1201. Society for Economic Dynamics.

Shapiro, J. R. and A. M. Williams (2012). The role of stereotype threats in undermining girls' and women's performance and interest in stem fields. *Sex Roles 66*(3-4), 175–183.

Shi, Y. (2018). The puzzle of missing female engineers: Academic preparation, ability beliefs, and preferences. *Economics of Education Review 64*, 129–143.

Shin, D. D., M. Lee, J. E. Ha, J. H. Park, H. S. Ahn, E. Son, Y. Chung, and M. Bong (2019). Science for all: Boosting the science motivation of elementary school students with utility value intervention. *Learning and Instruction 60*, 104–116.

Sorrenti, G., U. Zölitz, D. Ribeaud, and M. Eisner (2020). The Causal Impact of Socio-Emotional Skills Training on Educational Success. CESifo Working Paper Series 8197, CESifo.

Statistik Austria (2019). Schülerinnen und schüler 2018/19 nach detaillierten ausbildungsarten und geschlecht. https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bildung/schulen/schulbesuch/index.html.

Sutter, M. and D. Glätzle-Rützler (2015). Gender differences in the willingness to compete emerge early in life and persist. *Management Science 61*(10), 2339–2354.

Sutter, M., M. G. Kocher, D. Glätzle-Rützler, and S. T. Trautmann (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review 103*(1), 510–531.

Sutter, M., C. Zoller, and D. Glätzle-Rützler (2019). Economic behavior of children and adolescents–a first survey of experimental economics results. *European Economic Review 111*, 98–121.

Tai, R. H., C. Qi Liu, A. V. Maltese, and X. Fan (2006). Planning early for careers in science. *Science 312*(5777), 1143–1144.

Tungodden, J. and A. Willén (2023). When parents decide: Gender differences in competitiveness. *Journal of Political Economy (forthcoming)*.

UNESCO (2021). Unesco science report – the race against time for smarter development. https://unesdoc.unesco.org/ark:/48223/pf0000377433.

Van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness? *Journal of the European Economic Association 20*(4), 1595–1618.

Webber, D. A. (2014). The lifetime earnings premia of different majors: Correcting for selection based on cognitive, noncognitive, and unobserved factors. *Labour economics 28*, 14–23.

Wirtschaftskammer Oesterreich (2020). Lehrlinge in oesterreich 2019. http://wko.at/statistik/jahrbuch/lehrlinge19.pdf?_ga=2.204849048.949329663.1593348585-2085231723.1587021821.

Wiswall, M. and B. Zafar (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies 82*(2), 791–824.

Yeager, D. S., P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa, et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature 573*(7774), 364–369.

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources 48*(3), 545–595.

Appendix

A Additional figures and tables

FIGURE A.1
Data collection in class

FIGURE A.2
Histogram of *STEM interest* in the validation study by gender (N=339)

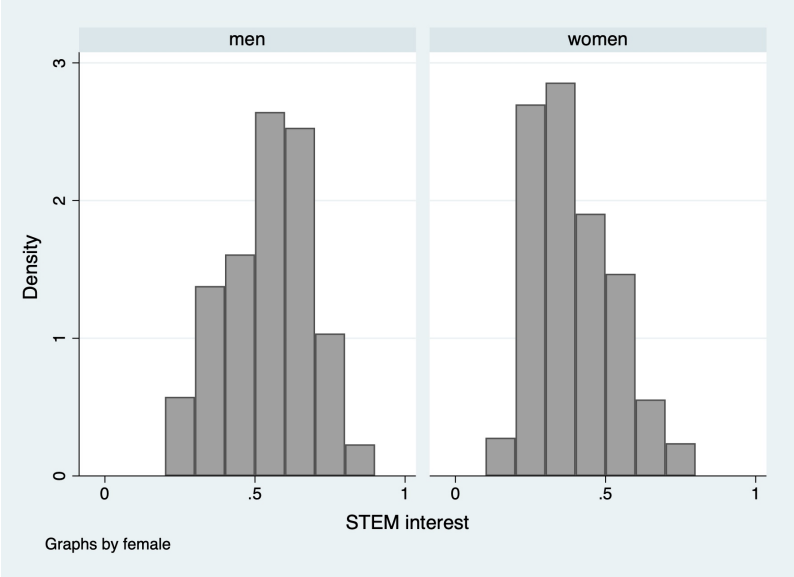


FIGURE A.3
Boxplot of *STEM interest* by *STEM occupation* in the validation study (N=345)

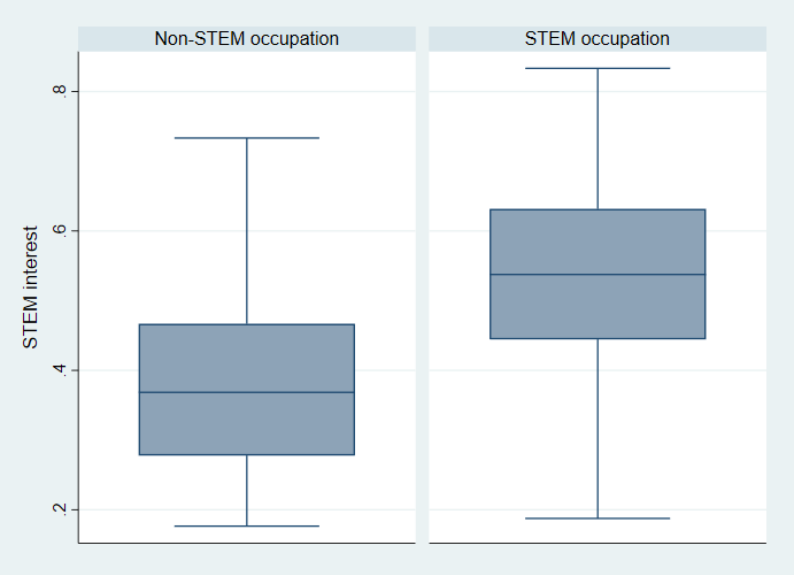


Table A.1
Logistic regressions to validate measure STEM interest

	Decision for STEM occupation					
	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled	Pooled	Women	Women	Men	Men
STEM interest	7.486***	7.451***	7.612***	7.521***	7.774***	7.488***
	(0.994)	(1.092)	(1.280)	(1.292)	(2.009)	(2.015)
Age		-0.280*		-0.277		-0.273
		(0.140)		(0.179)		(0.219)
Women		0.0874				
		(0.322)				
Vocational training		0.132		0.194		-0.0911
		(0.398)		(0.478)		(0.705)
Constant	-4.295***	-3.482***	-4.298***	-3.409***	-4.489***	-3.490**
	(0.488)	(0.771)	(0.579)	(0.799)	(1.136)	(1.326)
Observations	339	339	258	258	87	87
adj. R-squared	0.179	0.189	0.152	0.160	0.163	0.179

Note.— *Vocational training* is a dummy variable and 1 if graduate wants to pursue vocational training and 0 if s/he wants to enroll in a study program. Columns 1 and 2 present results of a logistic regression for the pooled sample, columns 3 and 4 for women and 5 and 6 for men. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table A.2
Treatment effects on mechanisms in the pooled sample

	(1) Explicit stereotypes	(2) IAT	(3) Growth mindset	(4) Compet.	(5) STEM Confidence
Treatment	-0.013**	-0.009	0.017*	0.031*	0.009
	(0.006)	(0.717)	(0.009)	(0.018)	(0.010)
Girls	-0.021***	-0.059	0.007	-0.096***	-0.147***
	(0.006)	(0.321)	(0.008)	(0.031)	(0.012)
Math performance				0.010**	
				(0.005)	
Baseline measurements					
Explicit stereotypes	0.071**				
	(0.033)				
Growth mindset			0.280***		
			(0.034)		
Competitiveness				0.398***	
				(0.033)	
Constant	0.483***	3.528*	0.384***	-0.224	0.655***
	(0.033)	(1.817)	(0.067)	(0.184)	(0.066)
Observations	962	962	962	962	962
R-squared	0.046	0.005	0.105	0.214	0.221
Controls	yes	yes	yes	yes	yes

Note.— Math ability represents the number of correctly solved tasks in the math task without competition. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors clustered by school. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A.3
Treatment effect on book choice (logit)

	Decision for a STEM book (in odds-ratios)					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.883	1.002	1.083	1.137	0.705	0.814
	(0.162)	(0.193)	(0.240)	(0.291)	(0.183)	(0.209)
Girls	0.697**	0.696**				
	(0.118)	(0.121)				
Constant	2.063***	3.667	1.283	2.797	2.355***	3.831
	(0.294)	(3.027)	(0.198)	(3.320)	(0.360)	(4.733)
Observations	962	962	480	480	482	482
Pseud. R^2	0.006	0.024	<0.001	0.024	0.005	0.045
Controls	no	yes	no	yes	no	yes
Gender differences		(3)-(5)	1.536		(4)-(6)	1.590
in treatment effects			(0.478)			(0.503)

Note.— Odds ratios based on logistic regressions with the choice of a STEM book as the dependent variable. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors clustered by school. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

TABLE A.4
TREATMENT EFFECTS ON INTEREST IN STEM CONTROLLING FOR BROCHURES

	STEM interest					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.028	0.038**	0.049**	0.074***	0.008	0.003
	(0.017)	(0.017)	(0.022)	(0.019)	(0.020)	(0.027)
Girls	-0.163***	-0.163***				
	(0.013)	(0.013)				
Brochures	0.011	0.012	0.008	0.022	0.016	0.005
	(0.018)	(0.018)	(0.021)	(0.017)	(0.026)	(0.030)
Treatment	-0.027	-0.023	-0.019	-0.021	-0.037	-0.030
× Brochures	(0.027)	(0.024)	(0.032)	(0.030)	(0.040)	(0.038)
Constant	0.588***	0.632***	0.413***	0.402***	0.600***	0.729***
	(0.013)	(0.065)	(0.013)	(0.104)	(0.014)	(0.066)
Observations	962	962	480	480	482	482
R-squared	0.207	0.222	0.017	0.046	0.004	0.051
Controls	no	yes	no	yes	no	yes
	Chooses a STEM book					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	-0.006	-0.014	0.042	-0.021	-0.052	-0.028
	(0.060)	(0.066)	(0.077)	(0.086)	(0.069)	(0.078)
Girls	-0.084**	-0.083**				
	(0.040)	(0.040)				
Brochures	-0.022	-0.067	-0.062	-0.132	0.021	-0.007
	(0.054)	(0.064)	(0.075)	(0.087)	(0.064)	(0.069)
Treatment	-0.052	-0.004	-0.055	0.046	-0.054	-0.047
× Brochures	(0.084)	(0.089)	(0.103)	(0.114)	(0.120)	(0.118)
Constant	0.685***	0.834***	0.594***	0.853***	0.692***	0.789***
	(0.041)	(0.192)	(0.055)	(0.280)	(0.045)	(0.266)
Observations	962	962	480	480	482	482
R-squared	0.012	0.035	0.010	0.041	0.008	0.059
Controls	no	yes	no	yes	no	yes

NOTE.— The upper panel shows OLS regressions with robust standard errors clustered at the school level and STEM interest as the dependent variable. The lower panel shows a linear probability model with robust standard errors clustered at the school level and the choice of a STEM book as the dependent variable. Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socio-economic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

TABLE A.5
DIRECT TREATMENT EFFECTS ON INTEREST IN STEM USING PARENTS REPORTED SES

	STEM interest					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.020	0.010	0.034	0.030	0.002	-0.014
	(0.016)	(0.017)	(0.021)	(0.023)	(0.023)	(0.026)
Girls	-0.179***	-0.167***				
	(0.015)	(0.016)				
Education	-0.008		-0.009		-0.007	
	(0.005)		(0.007)		(0.007)	
log HH inc.	-0.008		-0.009		-0.007	
	(0.005)		(0.007)		(0.007)	
Constant	0.671***	0.569***	0.522***	0.353*	0.640***	0.603***
	(0.096)	(0.122)	(0.121)	(0.175)	(0.151)	(0.151)
Observations	562	509	292	257	270	252
R-squared	0.261	0.237	0.036	0.035	0.057	0.045
Controls	yes	yes	yes	yes	yes	yes
Gender differences		(3)-(5)	0.007		(4)-(6)	0.015
in treatment effects			(0.029)			(0.030)
	Decision for a STEM book					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	-0.068	-0.030	-0.039	0.014	-0.110*	-0.090
	(0.050)	(0.055)	(0.067)	(0.070)	(0.064)	(0.074)
Girls	-0.095**	-0.111**				
	(0.042)	(0.046)				
Education	0.032*		0.005		0.056**	
	(0.016)		(0.021)		(0.022)	
log HH Inc.		0.006		0.008		0.003
		(0.022)		(0.041)		(0.022)
Constant	0.798**	0.921**	0.767*	0.777	0.761*	1.063**
	(0.339)	(0.432)	(0.420)	(0.591)	(0.382)	(0.489)
Observations	562	509	292	257	270	252
R-squared	0.053	0.048	0.033	0.031	0.119	0.087
Controls	yes	yes	yes	yes	yes	yes
Gender differences		(3)-(5)	0.076		(4)-(6)	0.120
in treatment effects			(0.081)			(0.086)

NOTE.— The upper panel shows OLS regressions with robust standard errors clustered at the school level and STEM interest as the dependent variable. The lower panel shows a linear probability model with robust standard errors clustered at the school level and choice of a STEM book as the dependent variable. Control variables are described in Table 1 and include children's age, spoken language at home, class specialization, and location of the school (urban or rural). All children participating in both data collections whose parents also answered the question concerning level of education (columns (1), (3), and (5)) or concerning household income (columns (2), (4), (6)) are included in the sample. Numbers in parentheses indicate standard errors.* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

TABLE A.6
DETERMINANTS OF STEM INTEREST BY GENDER (BASELINE MECHANISMS)

	(1) Girls	(2) Girls	(3) Girls	(4) Boys	(5) Boys	(6) Boys
Explicit stereotypes	0.035			0.068		
	(0.073)			(0.084)		
Growth mindset		-0.007			0.035	
		(0.054)			(0.048)	
Competitiveness			0.040**			0.006
			(0.016)			(0.016)
Constant	0.414***	0.439***	0.449***	0.695***	0.712***	0.731***
	(0.108)	(0.102)	(0.094)	(0.079)	(0.067)	(0.060)
Observations	480	480	480	482	482	482
R-squared	0.019	0.018	0.033	0.049	0.049	0.048
Controls	yes	yes	yes	yes	yes	yes
Gender differences	(1)-(4)	-0.054	(2)-(5)	-0.064	(3)-(6)	0.036
in treatment effects		(0.115)		(0.071)		(0.023)

NOTE.— OLS regressions with robust standard errors clustered at the school level. STEM interest is the dependent variable. Mechanism variables are measured at the baseline data collection. In columns (1)-(3) only girls are included in columns (4) - (6) only boys are included. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school, class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

B Attrition

As described in Section 3.5, we exclude 171 of 1133 observations (15%) from our main data analysis. To check if attrition may bias our results, we do the following. First, we investigate how attrition is distributed across schools, and, second, we test if attrition differs between treatment groups, gender, and school location (urban or rural). In general, attrition varies between schools. This is not surprising because, e.g., diseases spread within schools, and children stay at home for recovery. For a few schools, there are other explanations. In one school, it was not possible to conduct the first data collection due to technical problems. In another school, the different class-grade structure made it more demanding to have all children present for the data collection. There, the share of children who did not participate in both data collections is 42% higher than in the other schools (15%).

To test for biased attrition, we compare the ratio of girls, treated children, and children living in an urban area between the children included in the analysis (sample) and those who are not (dropouts). The share of treated children is 57% in our sample and 54% in the dropouts ($p = 0.599$, test of proportions), the share of girls in our sample is 50% and the share of girls in the dropouts is 49% ($p = 0.852$, test of proportions), and the share of children living in an urban area is 31% in our sample and 26% in the dropouts ($p = 0.264$, test of proportions). In Table B.1, we use logistic regression to confirm non-parametric results and test for joint significance. We do not find that any of the variables nor all of them together predict who participates in both data collections ($p = 0.927$, Chi-square test). Hence, we do not find evidence that there is biased attrition in our study.

Table B.1
Determinants of inclusion in the sample

	Probability to be in the sample
Treatment	1.110
	(0.438)
Girls	1.027
	(0.188)
Rural	1.243
	(0.548)
Constant	4.925***
	(1.355)
Observations	1,133
Pseud. R^2	0.002

Note.— Logistic regression with robust standard errors clustered at the school level and inclusion in the sample as the dependent variable. We present odds ratios in this table.

C Robustness checks

Decomposition of relative STEM interest measurement

TABLE C.1
TREATMENT EFFECTS ON AVERAGE INTEREST IN STEM AND NON-STEM JOBS

	AVERAGE INTEREST IN STEM JOBS					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.129	0.188**	0.300**	0.377***	-0.043	-0.011
	(0.102)	(0.090)	(0.140)	(0.118)	(0.097)	(0.102)
Girls	-0.551***	-0.545***				
	(0.075)	(0.072)				
Constant	2.635***	2.603***	1.987***	2.288***	2.732***	2.508***
	(0.072)	(0.312)	(0.084)	(0.571)	(0.069)	(0.587)
Observations	962	962	480	480	482	482
R-squared	0.080	0.121	0.021	0.101	0.001	0.030
Controls	no	yes	no	yes	no	yes
Gender differences		(3)-(5)	0.34**		(4)-(6)	0.33**
in treatment effects			(0.130)			(0.122)
	AVERAGE INTEREST IN NON-STEM JOBS					
	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
Treatment	0.053	0.065	0.037	0.020	0.068	0.132
	(0.100)	(0.089)	(0.087)	(0.092)	(0.147)	(0.139)
Girls	0.695***	0.693***				
	(0.069)	(0.066)				
Constant	1.913***	1.797***	2.616***	3.098***	1.904***	1.105**
	(0.083)	(0.360)	(0.041)	(0.518)	(0.102)	(0.508)
Observations	962	962	480	480	482	482
R-squared	0.117	0.162	0.000	0.040	0.001	0.082
Controls	no	yes	no	yes	no	yes
Gender differences		(3)-(5)	-0.03		(4)-(6)	-0.08
in treatment effects			(0.137)			(0.133)

NOTE.— Definitions of control variables are described in Table 1 and include children's age, spoken language at home, location of the school (urban or rural), class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

In the main paper, we use a composite measurement for STEM interest relative to interest in non-STEM jobs. While this relative measurement is relevant for career decisions (see validation study), it is still interesting to learn more about whether the treatment app increases interest in STEM jobs, as intended, or rather decreases interest in non-STEM jobs. In Table C.1, we replicate the upper panel of Table 2 using children's average interest in STEM jobs in the upper panel and children's average interest in non-STEM jobs in the lower panel. Both variables range from 0-3. We find that an increase in interest in STEM jobs drives the treatment effect reported in the main part of the paper. The average interest in STEM jobs increases significantly in the pooled sample when we include controls (column (2)) and a

dummy for girls (columns (3-4)). We find no evidence for a treatment effect on boys. Moreover, we find no evidence that the treatment affects the interest in non-STEM jobs (see lower panel).

Explicit stereotypes - Who can

In addition to the variables presented in the main text, we have elicited a second measurement for explicit stereotypes, implicit stereotypes, and confidence. We will discuss the robustness of our results with respect to using these alternative measurements in this section. Explicit stereotypes are measured by asking children who, men or women, are better at doing certain jobs. We call the variable *Who can* and it ranges from 0 anti-stereotypical views to 1 very stereotypical views. Table C.2 shows the manipulation check for *Who can*. We find that the treatment reduces stereotypical thinking for the pooled sample (column (1)) as well as for girls looking at the gender-disaggregated data (columns (2) and (3)).

Table C.2
Explicit Stereotypes - Who can

	(1) Pooled	(2) Girls	(3) Boys
Treatment	-0.017*	-0.018**	-0.014
	(0.009)	(0.008)	(0.015)
Girls	-0.042***		
	(0.007)		
Constant	0.594***	0.590***	0.580***
	(0.044)	(0.061)	(0.062)
Observations	962	480	482
R-squared	0.103	0.059	0.083
Controls	yes	yes	yes

Note.— OLS regressions with robust standard errors clustered at the school level. *Who can* is the dependent variable. Column (1) shows the aggregated results, and columns (2) and (3) show results for girls and boys separately. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school, class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Implicit stereotypes

As a second measurement for implicit stereotypes, we have asked children about who ranked in the top three in the math task. We compare their ranking of boys and girls with the actual ranking in class. The children earn points if their guess was correct. The variable *Implicit stereotypes* is an integer variable and ranges from -3 to +3 and is -3 when a child ranked only girls in the top three but only boys were among the top three performers. In contrast, when they ranked only boys in the top three, but the top three performers were only girls, the variable takes the value +3. The authors have developed this elicitation method and it relies on the assumption that children have some (biased) idea about who are the best performing children in the math task. Using a Poisson regression to identify determinants of the children's guessed rankings in the baseline measurement, we find that the number of boys or girls that are among the top 15 percentile of their class (with, on average, 20 children per class, this should represent the top three) does not affect the number of boys that children rank in the top three. To put it differently, actual performance is not informative for the children's guessed ranking.

Similar to the results for explicit stereotypical thinking, we find that implicit stereotypes are more pronounced for boys than for girls in the baseline measurement (average girls: -0.68 and average boys: 0.10, $p < 0.001$, Wilcoxon-Mann-Whitney test). Table C.3 presents the results of a manipulation check for implicit stereotypes. We find that the intervention did not reduce children's implicit stereotypes. In contrast, implicit stereotypes increase in the treatment compared to the control group. The increase in the pooled sample is driven by an increase for boys (see column (3)). A closer look shows that this effect is driven by a decrease in stereotypical thinking in the control rather than an increase in the treatment group. While the average ranked number of boys in the top three does not change for children in the treatment group (baseline: 1.93 vs. end: 1.92, $p = 0.813$, signtest), it decreases in the control group (baseline: 1.90 vs. end: 1.79, $p = 0.010$, signtest).

We test if the change might be explained by improvement in the math task. While children on average improve their performance between the two data collections (+0.43 tasks, $p < 0.001$, paired t-test), there is no gender difference in the improvement, neither in the treatment (average gender difference in improvement: +0.03 tasks, $p = 0.928$, two-sided t-test) nor in the control group (average gender difference in improvement: +0.11 tasks, $p = 0.670$, two-sided t-test).

TABLE C.3
IMPLICIT STEREOTYPES

	(1) Pooled	(2) Girls	(3) Boys
Treatment	0.462**	0.383	0.572***
	(0.204)	(0.237)	(0.199)
Girls	-0.532***		
	(0.115)		
Baseline	0.189**	0.234**	0.118
	(0.086)	(0.091)	(0.090)
Constant	-0.379	-1.081	-0.371
	(0.449)	(0.711)	(0.457)
Observations	924	454	470
R-squared	0.272	0.235	0.204
Controls	yes	yes	yes

NOTE.— OLS regressions with robust standard errors clustered at the school level. Implicit stereotypes are the dependent variable. We do not observe baseline implicit stereotypes for three classes due to technical problems during the baseline data collection. This reduces the sample to 924 observations. Column (1) shows the results for aggregated data, and columns (2) and (3) show results for girls and boys separately. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school, class specialization, and a proxy for socioeconomic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

There are several potential other explanations for this finding. First, the control intervention may decrease children's implicit stereotypical thinking. Second, our elicitation method was not successful to measure implicit stereotypes. We find no positive correlation between implicit stereotypes and explicit stereotypes or children's interest in STEM in the control group. Hence, explanation 1 seems unlikely. The second explanation is to some extent supported by our data as implicit stereotypes are not correlated with the results from the IAT, a validated measurement of implicit stereotypes. However, our analysis cannot provide a definite explanation for this finding.

Math confidence

In addition to the confidence in STEM ability, we also consider differences in confidence in the math task. We find that girls are, on average, less confident than boys in the baseline measurement. Boys overestimate their performance on average by seven tasks, and girls overestimate their math performance on average by five tasks ($p < 0.001$, Wilcoxon-Mann-Whitney test). Table C.4 shows a manipulation check for children's confidence in math. Columns (1) - (3) use a continuous measure, i.e., the difference between the guessed and the actual number of correctly solved tasks, and columns (4) - (6) show the share overstating their ability as registered in the pre-analysis plan. We do not find a significant treatment effect.

Table C.4
Math confidence

	Continuous variable			Share overestimating		
	(1) Pooled	(2) Girls	(3) Boys	(4) Pooled	(5) Girls	(6) Boys
Treatment	-0.497	-0.834	-0.213	-0.020	-0.021	-0.020
	(0.704)	(0.801)	(0.838)	(0.043)	(0.049)	(0.052)
Girls	-0.183			0.015		
	(0.423)			(0.025)		
Baseline	0.274***	0.346***	0.209***	0.211***	0.210***	0.211***
	(0.042)	(0.052)	(0.046)	(0.039)	(0.050)	(0.056)
Constant	-4.711	-1.842	-7.785**	0.356**	0.785***	0.017
	(3.225)	(4.740)	(3.613)	(0.169)	(0.248)	(0.214)
Observations	962	480	482	962	480	482
R-squared	0.144	0.197	0.125	0.065	0.082	0.069
Controls	yes	yes	yes	yes	yes	yes

Note.— OLS regressions with robust standard errors clustered at the school level. Math confidence is the dependent variable in columns (1) - (3), where we use the difference between guessed and actual numbers of correct answers in the math task, and in columns (4) - (6), the dependent variable is a binary variable with value one if the child overestimated his or her performance and zero otherwise. Columns (1) and (4) show the aggregated results, while we show results for girls and boys separately in the other columns. Control variables are described in Table 1 and include children's age, spoken language at home, location of the school, and a proxy for socio-economic status. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

D The effects of the brochure

In addition to the app, we distributed a brochure to the parents that explained why STEM skills are important for children and how they can support their children in acquiring them. The authors of the study were skeptical from the start regarding whether information brochures provide effective treatments since parents are often hard to reach, and it seemed unlikely that enough parents will receive, read, and fully grasp the details of the (German) brochure. However, public policy partners insisted on an intervention that addressed not only the children but also their parents. While we think that more specific and expensive treatments for parents could have effects, distributing brochures seemed unlikely to work well. We distributed the brochures to 50% of the participating schools. The brochure was distributed after the main intervention ended. To investigate the effect of the brochure on parents' and children's interest in STEM, we offer parents to participate in a lottery to win a voucher for summer camp in the end questionnaire. The summer camp offers four different themes: being creative, sports

activities, the city of Vienna, and STEM. We only consider children from Vienna as the summer camp was only available in Vienna. We measure the effect of the brochure by comparing the workshop choices between parents that did and did not receive the brochure. In total, 135 parents participated in the lottery. We find that parents of sons are more likely to sign up for the STEM workshop than parents of daughters. For boys, 46% of the parents chose the STEM theme; for girls, only 27% chose the STEM theme ($p = 0.019$, Fisher's exact test).

Concerning treatment effects, we find that 32% of parents in the control and 39% of parents in the treatment group sign their children up for a STEM workshop. This difference is not statistically significant based on a test of proportions ($p = 0.446$). The result does not change when including control variables in logistic regressions shown in Table D.1. We include controls for children's age, spoken language at home, and a proxy for socioeconomic status.[26] We do not find evidence that the brochure significantly affected the likelihood that parents signed up for the STEM theme (column (1)). The same holds when we consider boys and girls separately (columns (2) and (3)). However, our sample is reduced significantly, given the relatively low number of participants, not giving us the statistical power to detect an effect size of 0.33 standard deviations that we observe in our sample. We conclude that we do not find evidence that the brochure had a large effect (i.e., an effect of more than 0.60 standard deviations, which we would be powered to measure) on the share of children who were signed up for STEM-themed summer camp.

Table D.1
Decision for STEM workshop (odds ratios)

	(1) Pooled	(2) Girls	(3) Boys
Brochures	1.213	1.234	0.924
	(0.511)	(0.686)	(0.429)
Girls	0.423**		
	(0.142)		
Constant	2.368	0.269	6.160
	(8.240)	(0.955)	(32.429)
Observations	135	75	60
Pseud. R^2	0.044	0.027	0.024
Controls	Yes	Yes	Yes

Note.— Logistic regressions with robust standard errors clustered at the school level. The decision for a STEM-related workshop is the dependent variable and the estimates are reported as odds ratios. Column (1) shows the aggregated results, and columns (2) and (3) show results for girls and boys separately. Control variables are described in Table 1 and include children's age, spoken language at home, and a proxy for socioeconomic status. As only children in Vienna were able to join the workshop, we do not control for school location. Numbers in parentheses indicate standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

E Treatment effect on the treated

In line with the previous literature, we estimate the local average treatment effect (LATE) on the treated. While children from the control group could not receive the treatment by design, there are children in treated classes who did not use the app. To estimate the local treatment effect, we use treatment assignment as an instrument for app usage. Table E.1 shows the estimates for the LATE, i.e., the second

[26] We do not control for class specialization in these regressions, as the number of parents who signed up for a workshop is very low for some classes leading to perfect separability. As the workshop was only offered in Vienna, we do not control for the school location.

stage results of the regressions. As the take-up rate is 65% in the treatment group (see Table 1), the LATE estimator is approximately 35% larger than the intent-to-treat estimator reported in the main part of the paper.

The treatment was implemented at the class level. Therefore, we can also use a class-level definition of "treated" individuals. We define a class as treated if the children in this class used the app on average on more days than the app has been used on average across all treatment classes (see also our pre-analysis plan). Using this definition of treated classes, we run the LATE analysis again as described above in Table E.2. We find that 17% of the children in the treatment group are classified as treated under this definition. Consequently, the LATE is around 83% larger than the intent-to-treat estimator reported in the main part of the paper.

Table E.1
LATE on STEM interest (individual level)

	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled	Pooled	Girls	Girls	Boys	Boys
LATE	0.025	0.045**	0.061**	0.087***	-0.014	-0.015
	(0.023)	(0.021)	(0.025)	(0.026)	(0.030)	(0.033)
Constant	0.512***	0.524***	0.417***	0.442***	0.608***	0.741***
	(0.009)	(0.064)	(0.010)	(0.092)	(0.013)	(0.062)
Observations	962	962	480	480	482	482
Controls	no	yes	no	yes	no	yes

Table E.2
LATE on STEM interest (class level)

	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled	Pooled	Girls	Girls	Boys	Boys
LATE	0.115	0.148	0.241*	0.257**	-0.075	-0.057
	(0.113)	(0.096)	(0.144)	(0.125)	(0.166)	(0.126)
Constant	0.512***	0.532***	0.417***	0.457***	0.608***	0.738***
	(0.009)	(0.058)	(0.010)	(0.090)	(0.013)	(0.060)
Observations	962	962	480	480	482	482
Controls	no	yes	no	yes	no	yes

Below we estimate the LATE for the book choice. In Table E.3, we use treatment assignment as an instrument for app usage at the individual level. In Table E.4, we focus on treated classes as described above. While the estimates increase in size comparably to the results for STEM interest, we do not find a significant local average treatment effect under these model specifications.

Table E.3
LATE on STEM book (individual level)

	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
LATE	-0.044	0.003	0.030	0.045	-0.122	-0.072
	(0.066)	(0.067)	(0.081)	(0.087)	(0.093)	(0.090)
Constant	0.632***	0.742***	0.562***	0.756***	0.702***	0.812***
	(0.027)	(0.180)	(0.038)	(0.280)	(0.032)	(0.256)
Observations	962	962	480	480	482	482
Controls	no	yes	no	yes	no	yes

Table E.4
LATE on STEM book (class level)

	(1) Pooled	(2) Pooled	(3) Girls	(4) Girls	(5) Boys	(6) Boys
LATE	-0.200	0.011	0.117	0.134	-0.646	-0.276
	(0.326)	(0.223)	(0.323)	(0.261)	(0.614)	(0.379)
Constant	0.632***	0.743***	0.562***	0.764***	0.702***	0.799***
	(0.027)	(0.180)	(0.038)	(0.280)	(0.032)	(0.267)
Observations	962	962	480	480	482	482
Controls	no	yes	no	yes	no	yes