Econometric Analysis of the Effects of Educational Decisions on Labor Market Outcomes and the Influence of Self-Testing on Learning Outcomes

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts - und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Jakob Schwerter

aus Iserlohn

Tübingen

2020

Tag der mündlichen Prüfung:	07.12.2020
Dekan:	Professor Dr. rer. soc. Josef Schmid
Erstberichterstatter:	Professor Dr. rer. pol. Martin Biewen
Zweitberichterstatter:	Professor Dr. rer. pol. Joachim Grammig

Publication notes

Chapter 2 is based on Ilg, L. and J. Schwerter (2020): Gender differences in the labor market entry of STEM graduates: Does fertility play a role?, unpublished manuscript, University of Tübingen.

Chapter 3 is based on Biewen, M. and J. Schwerter (2020): Does more math in high school increase the share of female STEM workers? Evidence from a curriculum reform, unpublished manuscript, University of Tübingen.

Chapter 4 is based on Schwerter, J. (2020): Impact of universities in a flat hierarchy: Do degrees from top universities lead to a higher wage?, unpublished manuscript, University of Tübingen.

Chapter 5 is based on Schwerter, J., J. Bleher, T. Dimpfl and K. Murayama (2020): Practice makes perfect? Self-testing with external rewards, unpublished manuscript, University of Tübingen and University of Reading.

Chapter 6 is based on Schwerter, J.(2020): Practice makes perfect? Evidence from a voluntary self-testing setting, unpublished manuscript, University of Tübingen.

Chapter 7 is based on Schwerter, J., F. Wortha, P. Gerjets (2020): The added value of hints in multiple-try feedback: Can feedback enhance students' achievement during the semester?, unpublished manuscript, University of Tübingen,

Funding

This research was supported as part of the LEAD Graduate School & Research Network [GSC1028], which was funded within the framework of the Excellence Initiative of the German federal and state governments until October 31, 2019. Thereafter, it was funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II.

Danksagung

Die vorliegende Dissertation ist über vier Jahre im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Statistik, Ökonometrie und Quantitative Methoden an der Eberhard-Karls-Universität Tübingen entstanden. Davor und während dieser Zeit habe ich Unterstützung von vielen Seiten erfahren und möchte deshalb an dieser Stelle einigen Menschen danken, ohne die das Gelingen dieser Arbeit nicht vorstellbar gewesen wäre.

Als Erstes möchte ich meinen Weggefährten im Bachelor in Mannheim danken. Ohne diese hätte ich vielleicht nicht genug Spaß am Studium gehabt und meine Freude an der Datenanalyse wäre mir nicht bewusst geworden. Viele Kommilitonen aus der Zeit könnte ich nennen, aber vor allem die zusammen durchlebte Zeit während der Bachelorarbeit mit Franziska Ramschütz und Dennis Köhn bleiben für immer in Erinnerung. Ohne euch hätte ich über meine Fehler und Probleme nicht lachen können. Besonders hervorheben möchte ich weiter meinen Bachelorarbeitsbetreuer Stephan Kastoryano. Danke auch für die Gespräche während meines ersten, nicht so erfolgreichen Masterstudiums in Konstanz, die mich bestärkt hatten den Master dort abzubrechen und einen neuen in Tilburg zu starten. In Tilburg hat dann vor allem die Masterarbeit unter meinem Betreuer Martin Salm so viel Freude bereitet, dass ich wusste akademisch bleiben zu wollen.

Als Nächstes möchte ich mich dann bei Martin Biewen bedanken, der aus dem Wollen ein Dürfen gemacht hat. Vielen Dank für die Betreuung, Unterstützung, das Vertrauen und den Austausch als Doktorvater und Co-Autor. Ebenso möchte ich mich bei meinem Zweitbetreuer Joachim Grammig für die Unterstützung, das Feedback und jede Spaß-Wette (natürlich schaffte es Schalke nicht in die CL...) während meiner Zeit in Tübingen bedanken, besonders ab der Gründung der "Statistik-WG“. Zur allgemeinen Unterstützung in schwierigen Zeiten während der Promotion sowie zu vielen schönen Momenten haben natürlich auch alle weiteren "Bewohner“ beigetragen: Lea Eiting, Sylvia Bürger, Matthias Seckler, Madalina Thiele, Marian Rümmele, Thomas Dimpfl, Jantje Sönksen, Johannes Bleher, Dalia Elshiaty, Constantin Hanenberg und mein „Roomy“ Miriam Sturm. Vielen Dank für alle Pausen, fachlichen, allgemeinen und spaßigen Diskussionen, Weihnachtsfeiern, Kuchen, Biere, Sektflaschen und aufbauende Ratschläge und Meinungen. Ohne euch wäre die Zeit mindestens deutlich schwerer und einsamer gewesen. Einen weiteren Dank an Miriam und Marian für die Zusammenarbeit an unserem Paper. Ohne euch wäre das Paper nicht publiziert worden (#effizient #EJPE). Einen besonderen Dank geht noch an meine weiteren Co-Autoren Franz Wortha, Johannes Bleher und Thomas Dimpfl für die super Mitarbeit an den Studien, ohne die ich nicht rechtzeitig fertig geworden wäre! An Lea Eiting geht noch der größte Dank für die gesamte Hilfe im administrativen Bereich und den Erfahrungswerten der Statistik Vorlesungen. Ohne dich wäre ich in einigen Situationen im Chaos untergegangen. Einen großen Dank auch an alle Hiwis, Tutoren und Tutorinnen für die Unterstützung in der Forschung und der Lehre.

Weiter möchte ich mich allgemein bei LEAD bedanken, vor allem bei Ulrich Trautwein, Benjamin Nagengast, Nicolas Hübner, Patricia Goldberg, Franz Wortha, Sophie Freitag und Mareike Bierlich. Durch LEAD habe ich eine zusätzliche fachliche Unterstützung bekommen und die Möglichkeit, meine Forschung mit einem großen Netzwerk zu teilen und Feedback zu kriegen. Weiter haben die Retreats und andere soziale Events meine Zeit in Tübingen bereichert.

Ein Dank geht natürlich auch an Freunde und Familie, die mich bis hierhin (oder zeitweise) begleitet haben. Ohne euch wäre ich anders, hätte andere Lebensentscheidungen getroffen, wie z.B. Politik und nicht VWL zu studieren (shout-out an Frederik) und wäre nicht in die Situation gekommen eine Danksagung schreiben zu dürfen

Ein letzter Dank geht an Stefan Glaisner, Miriam Sturm und Ines Pieronczyk fürs fleißige Korrekturlesen (#Legastheniker).

Jakob Schwerter, Tübingen im September 2020

Contents

List of Figures

List of Tables

Chapter 1

Dissertation Introduction

Despite considerable differences of opinions worldwide, educating the young seems to be one of few consensuses. The UNESCO, for example, reports a global rise in literacy rates for the last 40 years (UNESCO, 2017). The World Bank is one of many institutes promoting easier access to education (for the poor regions around the world) as well as an increase of the general quality of schooling (World Bank, 2011). One main reason for the strong emphasis on education is its positive impact on economic growth, innovations for the economy, employment perspectives, higher wages, and improved health for the individual as well as on reduced crime and poverty (OECD, 2017). The academic literature also regards education as essential for the future due to its high returns on both at the individual and at the aggregate level (Acemoglu & Angrist, 2000; Card, 1999; Grossman, 2006). The returns to education are empirically found for an individual's income, occupation status, individuals' health, social interactions, and cognitive and noncognitive skills (Card, 1999; Dickson & Harmon, 2011; Grossman, 2006). The classic model to test these returns to education is the Mincer equation (Mincer, 1974):

$$ln(y_i) = \alpha + \beta S_i + \gamma E_i + \rho E_i^2 + \epsilon_i \ , \tag{1.1}$$

where $ln(y_i)$ is, for example, the natural logarithm of the individuals' earnings, S_i the years of schooling an individual i completed, and E_i the amount of experience an individual has in the labor market. E^2 is included to take into account possible nonlinear effects of experience. One common change of this equation is the replacement of years of schooling by indicators for the highest degree an individual has obtained, known as the *sheepskin effect*. This effect allows education to have a non-parametric effect and, thus, not restricted to a polynomial function.

The effect of education in Equation (1.1) can be twofold: on the one hand, education may lead to better skills that increase individual productivity (i.e., human capital accumulation). That means, skills learned and improved in school lead to maximization of the potential work outcome like accelerated wages and job positions. Further, social skills may enable the individual to enjoy her leisure time, i.e., being healthy and more satisfied in life. On the other hand, schooling may (also) serve as a signal for potential employers that one is qualified for more prestigious work. “If those with more schooling also have more inherent abilities, employers can use schooling to predict better candidates. This is especially helpful when desirable worker attributes, like perseverance, discipline, and time management, are not easily observed” (Oreopoulos & Salvanes, 2011). Due to unknown information, a good signal can help companies distinguish qualified job candidates from others. Therefore, signaling eases employers’ and employees’ matching processes, especially helping individuals find the ‘right’ place and reduce their search costs. In case education does not only reveal inherent abilities but further helps develop and increase skills and the respective potentials, education adds to inherent abilities. Then, returns to education for the individual are only a lower bound estimate for the returns to education to the society, because increased potential may improve the entire economy (Oreopoulos & Salvanes, 2011).

Grossman (2006) names two possible models to explain how to gain returns to education on the individual level. He starts with the *productive efficiency* model: improvements in skills via schooling increase productivity, resulting, for instance, in higher wages. An example of this may be the computer scientist whose code becomes more efficient and may even make some employees and their work obsolete. As a second model, Grossman (2006) names the *allocative efficiency* model. This model focuses on improved skills contributing to a more efficient use of resources, for example, the reduction of costs. More educated individuals are then either more resource-efficient or less costly. Following Malamud (2011), at least at the university level, education matches individual qualities and job required qualities rather than increasing specific individual skills. While these models will not be tested in this dissertation, they form the individual chapters’ basis.

What did the literature find so far on the returns to education? An almost unaccountable (Card, 1999) amount of studies show that more educated individuals receive higher wages, are less likely to be unemployed and are more likely to work in more prestigious jobs than less-educated counterparts. For example, Angrist and Krueger (1991), Ashenfelter and Krueger (1994), Card (1995), Ashenfelter and Rouse (1998), Angrist and Evans (1999), Oreopoulos (2007), Oreopoulos and Salvanes (2011), Henderson, Polachek and Wang (2011), Machin, Salvanes and Pelkonen (2012), and Powdthavee, Lekfuangfu and Wooden (2015) reveal a causal link between education and income as well as the employment status of individuals. Further, education also leads to a higher job prestige (for example, Oreopoulos & Salvanes, 2011; Zhou, Lin & Lin, 2016) and job satisfaction (for example, Winkelmann & Winkelmann, 1998; Zhou et al., 2016). Oreopoulos (2007) and Machin et al. (2012) further show that individuals with more years of schooling are more likely to move to new cities, states or countries. This is important for countries where unemployment is also caused by missing mobility (Machin et al., 2012).

Education does not only increase skills, followed by higher productivity and higher wages but also improves non-pecuniary outcomes. Oreopoulos (2007), Oreopoulos and Salvanes (2011), Zhong (2015), and Zhou et al. (2016) find that schooling, aside from job satisfaction, increases also overall life satisfaction. Moreover, higher levels of education seem to improve individuals' health. The positive impacts of schooling on self-reported health are shown, among others, by Kemptner, Jürges and Reinhold (2011), Oreopoulos and Salvanes (2011), Powdthavee et al. (2015), Zhong (2015), and Zhou et al. (2016). Oreopoulos and Salvanes (2011) further illustrate that more education reduces the probability of a stay in a mental institution. Further, de Walque (2010), and Oreopoulos and Salvanes (2011) show that more educated individuals are more likely to stay alive within the next 10 years. Additionally, more educated individuals tend to profit earlier than others from newly found cures (Grossman, 2006).

Lochner (2004), Lochner and Moretti (2004), and Hjalmarsson, Holmlund and Lindquist (2015), among others, explore the relationship between educational investments, work, and crime. They show that an increase in skills and wages lead to a higher cost of unskilled crime. However, white-collar crimes can increase with skills but are negatively related to income.

Following Milligan, Moretti and Oreopoulos (2004), "Economists, educators and politicians commonly argue that one of the benefits of education is that a more educated electorate enhances the quality of democracy". Dee (2004) indicates an increase in civic participation by the amount of newspaper readership, group membership, and the acceptance of allowing a minority to speak freely due to more years of schooling. Oreopoulos and Salvanes (2011) show that more educated individuals were more likely to have voted in the last election. Additionally, Helliwell and Putnam (2007) provide evidence about civic participation benefits through schooling. Further, Helliwell and Putnam (2007) state that education is one of the most important predictors of trust due to "relative reasons (schooling raises social status), additive reasons (schooling teaches students how to interact properly) or superadditive reasons (schooling increases education attainment levels which makes everyone more trusting)." They find a positive effect on trust and social engagement for more educated individuals, but also for average education.

For women, more education is linked to more competitive behavior on the marriage market (Lafortune, 2013). Jones, Schoonbroodt and Tertilt (2010) underline that more educated women have fewer children and are more likely to work. These effects might result in more assortative mating, yielding an increased income inequality on the macroeconomic level, at least for the US, as shown by Greenwood, Guner, Kocharkov and Santos (2014).

Most of the results, however, have been established for the majority of the respective population. Oreopoulos and Salvanes (2011) highlight that no inference can be drawn from overall analysis to that of subgroups. A few exceptions concentrate on the differences between the majority and minorities. For example, taking into account that females are still somewhat of a minority in the labor market compared to males, Belman and Heywood (1991) find that the sheepskin effect for increased wages is higher for women with a higher level of education than for men with comparable education. Walker and Zhu (2011) confirm this tendency: women tend to benefit more from tertiary education compared to males.

Lastly, increased investment in human capital also impacts the macroeconomic growth of a respective country. Education can shape society by social and civic returns to education due to superadditive effects (Pritchett, 2006). This relationship is further shown in Gylfason (2001), Goldin and Katz (2008) and Barro (2015).

In the developed world, the individual decides to receive more education after finishing the compulsory school years. As shown by the OECD (2016), more and more citizens decide to go for tertiary education. But does achieving this level of education result in the same employment possibilities and wage levels for all graduates? Put differently: what does the individual's decision for a certain education level mean for the employment possibilities and wage level of the respective graduates?

Following Grossman (2006), education should increase productivity or make the use of resources more efficient, leading to the improved matching of individuals' skills and job requirements. This should be especially the case for higher education graduates (Malamud, 2011). If this is true, graduates should find a degree-related occupation due to the better matching, and not just "a" job. Therefore, the question arises whether tertiary education does not only lead to higher probabilities of employment but also to a higher probability of ending up in occupations that are related to the particular education an individual received. In Chapter 2 in combined work with Lena Ilg, I analyze if women opting to graduate in a field of study in the areas of science, technology, engineering, and mathematics (STEM) also stay in the field by working in a degree-related occupation. Does the decision to study STEM also lead to a STEM occupation? The analysis shows that female STEM graduates are, compared to their male counterparts, less likely to work degree-related.

Related to this, in combined work with Prof. Martin Biewen, we look at a curriculum reform of increased levels and hours of mathematics and natural sciences classes at the last two years of high school and analyze if such measures might help to increase the share of females working in STEM. In other words, can politicians reshape schools to influence high school graduates willingness to enter and stay in the STEM pipeline? According to our results, this is not the case.

Looking at the Mincer equation, schooling is considered mostly at the extensive margins: years of schooling or type of degree and alike. Education is, however, known to be heterogeneous, and one should also be concerned about the intensive margins. When individuals reach the same level of education, does it matter which institute they went through? Focussing on university graduates, I analyze whether the decision to study at a top ranked university increases the wages of these graduates. This question has already been addressed, especially for the USA, England, Australia, and some other countries, but not for Germany - a country with a rather flat university hierarchy. Thereby, I examine the heterogeneity in human capital accumulation and signaling effects discussed above.

To receive proper education seems to be an improvement for the respective individuals and society overall. Therefore, in the second part of the dissertation, I want to focus on how we might be able to help students attain knowledge and competences. How do students learn, and can we improve students' achievement outcomes? Is there an immediate return of more intensive studying on learning outcomes? With the availability of the internet and computers, e-learning environments become a prominent tool in (higher) education. During the last twenty years, university teaching saw a significant increase in the use of e-learning environments, especially in the course of the COVID-19 pandemic in 2020, where such tools, in fact, constituted the only option of teaching in many countries. If higher education uses these environments, it should be known how students are optimally helped to learn and enhance their learning outcomes. Simultaneously with these developments, research on these tools arises. The academic literature in the fields of economics, education, and psychology, which evaluates this rather new way of teaching has been growing accordingly.

The literature usually compares four (or less) types of learning environments: (i) face-to-face or live (classroom) teaching, which is what we were mostly used in the past, (ii) e-learning or online teaching using only recorded lectures or videos and online exercises with no face-to-face interactions (only optional online-meetings), (iii) hybrids of the two, which are called blended learning,[1] and (iv) the flipped classroom, a specific case of the blended teaching. The flipped classroom characteristics that the lecture is recorded and students have to work through the materials provided online before coming to a face-to-face meeting. During these face-to-face meetings, the (academic) teacher should be

[1] Review of blended teaching studies: M. G. Brown (2016)

available to help students apply the topics taught in the recorded videos and answer their questions concerning the provided worksheets and online tasks.

Numerous studies find that, even though one speaks of the great potential of online learning, that students being taught face-to-face usually outperform students taught online (for example, Alpert, Couch, Harmon & R, 2016; Bettinger, Fox, Loeb & Taylor, 2017; Bowen, Chingos, Lack & Nygren, 2014; B. W. Brown & Liedholm, 2002; D. Coates, Humphreys, Kane & Vachris, 2004; Figlio, Rush & Yin, 2013; D. Xu & Jaggars, 2014; Y. J. Xu, 2013). B. W. Brown and Liedholm (2002), and Alpert et al. (2016) analyze differences between a face-to-face, blended and virtual classroom in first-semester principal microeconomics. In B. W. Brown and Liedholm (2002), students who participated in the virtual classes seem to have "better" characteristics but worse exam grades, compared to the face-to-face taught students. The authors conclude from this that the virtual classroom needs more discipline to master the course. Further, there was no significant difference for the students taught according to the blended method compared to both the "face-to-face" and "online" students. D. Coates et al. (2004) confirm the results and add, however, that students who selected themselves into online classes benefit from the videos.

For blended teaching formats, Bowen et al. (2014) find in a randomized setting that face-to-face and blended formats do not lead to differences in exam grades (but underline the advantage of the blended format of saving money). Joyce, Crockett, Jaeger, Altindag and O'Connell (2015) only find a difference between face-to-face and blended formats in midterms but not in the final exam. The driving force of these results are students of the lower percentiles. Good students perform well in both settings, while low-ability students tend to suffer in the blended format. Fischer, Baker, Li, Orona and Warschauer (2019) found one positive result for virtual learning: even if students have lower achievements in online classes, they might still finish their studies in time with a higher probability.

O'Flaherty and Phillips (2015) review flipped classroom literature and summed up that flipped classrooms have the capacity for building lifelong skills for 21st century learners. Thai, De Wever and Valcke (2017) compare the flipped classroom, blended learning, face-to-face learning and online learning. They show that the flipped classroom results in higher learning performance compared to face-to-face and fully e-learning settings.

As shown in the flipped classroom study by O'Flaherty and Phillips (2015), video-lecturing seems to have great potential because students can watch the videos at their own time and pace, or even rewatch them. However, an important aspect that should not be forgotten when we think about changing teaching methods is noted in H. Coates (2006) and Barkley (2010): student engagement is essential for effective teaching and learning. Thus, while expanding courses with e-learning components, student interactions with the higher education teachers are important to be maintained. This is also in line with Jaggars and Xu (2016). Their results indicate that the quality of interpersonal interaction within a course relates positively and significantly to student grades. Additional analyzes based on course observation and interview data suggest that frequent and effective student-instructor interaction creates an online environment that encourages students to commit themselves to the course and perform better.

Broadbent and Poon (2015) additionally find for online courses that self-regulated learning strategy effects are weaker in the online context than in the traditional classroom. Following their results, the success of online learning is based on the prioritization of peer learning. Further, Jaggars (2014) show that students favor fully online learning for easy subjects but face-to-face lectures for more difficult and important subjects.

The mentioned results in this upper paragraph show possible reasons why face-to-face learning outperforms online learning so far. Flipped classrooms surpasses both formats, because it features the best of both worlds. Students can work on their own pace while having the interaction with other students as well as the teaching staff.

While most of the studies focus on the broad differences in learning outcomes between face-to-face, video lecturing, and a mixture of the two, the studies lack the specific analysis of the effects of increased self-testing possibilities in e-learning environments. Fischer, Zhou, Rodriguez, Warschauer and King (2019) is one exception that shows that a three-week preparatory online course can help improve students' performance in a course in the following semester. The e-learning studies included in this dissertation focus on self-testing possibilities in e-learning environments within the semester. In other words, two studies assess whether students who are allowed to practice with direct feedback get better grades. Moreover, one study examines whether we can enhance students' performance by giving additional explanation after an incorrect response. Therefore, in the second part of the dissertation, I focus on the potential of the practice possibilities of e-learning environments and how they can help to improve the acquisition of knowledge and learning outcomes such as grades at the end of the semester.

Once together with Johannes Bleher, Thomas Dimpfl, and Kou Murayama, and once in a study by myself, I analyze whether practicing is beneficial for students. In the first study, we analyze the self-testing impact in a setting with three voluntary midterms and additional practice possibilities. Good performances in the midterms lead to extra points added to the final exam points. In the second study, the self-testing is still voluntary, but weekly and without additional rewards. Though we cannot claim a causal link, we, at least, show for two data sets and two different settings that self-testing during the semester predict higher exam scores. In both studies, we control for a very rich set of control variables like ability, motivation, and personal traits.

In addition to these two studies, together with Franz Wortha and Peter Gerjets, I raise the question of whether we can increase students' learning gains with additional hints in an e-learning sessions. Therefore, we use a setting of mandatory e-learning participation and a within-semester-randomization to see whether additional hints to the knowledge of correct response helps students to perform better when solving the e-learning exercises. These three studies provide evidence for the effectiveness of an online teaching method for statistics or mathematics at higher education institutions.

To conclude this introduction, I want to give an overview of the six studies this doctoral thesis comprises. The first three studies focus on the effect of educational decisions on labor market outcomes, while the second part concentrates on the influence of self-testing on learning outcomes.

Gender differences in the labor market entry of STEM graduates:
Does fertility play a role?

Chapter 2 of the thesis is motivated by the result of Malamud (2011) mentioned above: higher education should help to match occupational requirements and individual skills. But does everyone work in a degree-related occupation after graduating from university? This is questionable in some places. There is a public and academic debate whether women, even if they decided to study a STEM subject, are less likely to work in a STEM occupation. More specifically, Lena Ilg and I analyze if female STEM graduates are equally likely than men to work in a degree-related occupation after graduation or not. Suppose men and women studied, for example, STEM, to match individuals' skills and interests with job requirements. In that case, both genders should be similarly likely to work in a degree-related occupation. We show that children born before graduation do not contribute to the over-proportionate non-transition of women to STEM occupations. The educational decision is personal, i.e., individuals decide themselves which field of study they choose. The labor market outcome is the likelihood to work degree-related.

Does more math in high school increase the share of female STEM workers?
Evidence from a curriculum reform.

We know from Chapter 2 that female STEM graduates are less likely to have a degree-related occupation. In Chapter 3, Prof. Dr. Martin Biewen and I investigate if increased math requirements at the end of high school in one of the German federal states increased the share of male and female students who complete degrees in STEM subjects and who later work in STEM occupations. The reform had two important aspects: (i) it equalized all students' exposure to math by making advanced math compulsory in the last two years of high school; and (ii) it increased the instruction time from three to four hours per week and raised the level of instruction in math and the natural sciences for some 80% of the students, more so for females than for males. Our results suggest that, despite its substantial nature, the reform did not change the share of men completing STEM degrees and that it even reduced the share of women graduating from STEM programs. Moreover, we do not find general reform effects on the share of individuals working in STEM occupations after graduation for both men and women. The education decision here is made at the political level and is thereby obligatory for all individuals within the reform state. The effects on labor market possibilities are the probability of working in STEM or of one subgroup within STEM.

Impact of universities in a flat hierarchy:
Do degrees from top universities lead to a higher wage?

For Chapter 4 of my dissertation, I analyze the importance of the decision at which university individuals study and graduate. This chapter does not directly build on the last and is motivated by the Mincer Equation (1.1) with the variation that I do not look at different lengths of education but being educated by more prestigious universities. Thus, the analysis is about the intensive margin of education. The general literature shows a wage premium for graduates from high quality, elite, or more selective universities. These results, however, have been established for countries with a clear hierarchy of top universities, such as the US, England, and Australia. I evaluate if such an effect also exists in Germany, a country where individual universities are top-performing in some but not necessarily all fields. Further, the general differences between universities are smaller compared to, for example, the US. I use the University Ranking of the Quacquarelli Symonds and a revealed preference and acceptance ranking to measure a university's quality. Both rankings show a wage premium in IV regression in-between 5 and 13%. This effect is especially pronounced for women, relating to the results in Belman and Heywood (1991), and Walker and Zhu (2011), which claim that women tend to profit more from additional education.

Practice makes perfect?
Self-testing with external rewards.

In Chapter 5, Johannes Bleher, Dr. Thomas Dimpfl and Prof. Dr. Kou Murayama observe students of the course *Mathematical Methods in Economics and Business Administration*. These students were offered three midterms in the e-learning environment and allowed to self-test themselves afterward without additional rewards. Further, we included an application in which students could playfully test their knowledge in one specific topic and see how they rank within all students who use this application. Here, again, we find that participating in the midterm (and the additional test runs) is beneficial for the exam. Performance in the midterms and further practices are additionally beneficial. For the application, we also find a performance effect, but no submission effects.

Practice makes perfect?
Evidence from a voluntary self-testing e-learning setting.

In Chapter 6, I analyze if voluntary practice in a setting without external rewards can help students to achieve better grades and if we can relate certain characteristics of individuals to participate in the voluntary self-testing. To control for selection into participation, I control, among others, for important predictors for educational success, namely ability, motivation, personality traits, and goals. To answer the research question, we observe sociology students working voluntarily on (weekly) online-exercises. I find that, even after controlling for variables as mentioned earlier, participation leads to an increase of points in the end exam. Interestingly, the performance itself was not crucial in this setting.

The added value of hints in multiple-try-feedback:
Can feedback enhance students' achievement during the semester?

Chapter 7 addresses the question of whether we can increase students learning outcomes with additional feedback in e-learning exercises. Therefore, with Franz Wortha and Prof. Dr. peter Gerjets, we compare two groups: students who, while solving exercises, get only feedback of correct response and a group that received additional hints after answering the exercise. We conduct this experiment with third-semester sociology students in the tutorial for the class *Social Science Statistics II*. We show that the additional feedback helps students perform better within the session and in a former exam question the week after.

Chapter 2

Gender differences in the labor market entry of STEM graduates: Does fertility play a role?*

*This chapter is based on: Ilg, L. and J. Schwerter (2020): Gender differences in the labor market entry of STEM graduates: Does fertility play a role?, unpublished manuscript, University of Tübingen.

2.1 Introduction

The under-representation of women in sciences, technology, engineering, and mathematics (STEM) fields of higher education programs and occupations has received considerable attention in the scientific literature and public debate in recent years. Almost all of the STEM occupations continue to be dominated by men. At the same time, many women do not enter STEM occupations even though they have graduated with a STEM degree. Recent numbers illustrate this phenomenon: in 2016, women accounted for 28% of STEM graduates in Germany. At the same time, only around 19% of the STEM workforce was female. In contrast, women represented nearly half of all university graduates (48.5%), as well as almost half of the entire workforce (46%) in Germany in 2016 (Bundesagentur für Arbeit, 2019). Hence, women are still underrepresented in the STEM workforce and seem to have difficulties entering it.

At the same time, women's contribution to STEM occupations is considered crucial for the innovative power and the continuous development of the STEM sector. Due to population aging, a male-dominated workforce, and to meet future demands, the industry needs female workers in STEM (Burke, 2007). Scholars also stress that women bring in new thinking styles and different approaches to problem-solving, which may lead to production increase (Simard, Henderson, Gilmartin, Schiebinger & Whitney, 2013). Since the STEM sector is crucial for a country's economic success, women not entering STEM occupations are not only important from a gender equality perspective. It can also substantially impact companies' economic performance because of second-tier men taking the place of women who would have been better prepared but who leave STEM (Justman & Mendez, 2018). The underlying reasons for the underrepresentation of women in this area remain mostly unknown. It is, however, well-documented that STEM occupations are still overly hostile to female workers (for example, Danbold & Huo, 2017; Simard et al., 2013).

The importance of the issue is also reflected in a number of studies in the empirical literature. A considerable part of these studies consists of exit studies that focus on the retention of women (for example, Hunt, 2016; Kahn & Ginther, 2015; L. A. Morgan, 2000; Preston, 1994). These studies find univocally higher exit rates for women compared to men within STEM. Thereby, only Hunt (2016) includes non-STEM graduates and workers to check if the effect is general for women or field-specific. However, the exit studies do not distinguish between a missing entry or an exit during the career. From a policy perspective, it is crucial to know whether women never entered or leave the STEM workforce. Only Sassler, Glass, Levitte and Michelmore (2017) provide a notable exception of studies that examine the gender differences in the transition to the first occupation after university. They highlight, however, only differences between the STEM subfields. Thus, we are, to the best of our knowledge, the first to focus on the transition phase from university to the labor market including both STEM and non-STEM graduates. Including both gender and STEM as well as non-STEM individuals enable us to show that a gender difference in transition behavior exists specifically in STEM occupations. Therefore, we consider it to be of great importance to include male and female graduates from both STEM and non-STEM fields of study in our analysis, similar to Hunt (2016).

One possible explanation of the gender difference are childcare obligations. Existing findings are mixed in this regard. Kahn and Ginther (2015) find that women with children are less likely to enter or stay in science and engineering careers. On the contrary, family expectations can not explain the different transition or remaining rates in STEM and non-STEM occupations in the studies of Preston (1994), Hunt (2016), and Sassler et al. (2017). Thus, our findings add evidence to this unanswered question.

Based on two waves of the cohorts 2005 and 2009 of the Graduate Panel from the German Centre for Higher Education Research and Science Studies (DZHW), the following analysis shows a field-specific advantage for males graduating in STEM to work in a degree-related occupation of around 6 to 7%. Then, interacting the study-field with the gender of the individuals shows a negative field-specific gender difference for women of around 4 to 5%. Thus, if women have a STEM-degree, there are less likely to work degree-related than their male counterparts. Comparing female STEM-graduates with non-STEM graduates shows that female STEM-graduates are more likely to work degree-unrelated. The results for the STEM-graduates are driven by the group of engineers and computer scientists (EngComp), which is why we focus on this group in our main analysis. For them, we find a slightly higher and significant field-specific gender difference working degree-related of around 5 to 6 percentage points. We do not find evidence that the relative higher non-entry rate of women in EngComp can be attributed to children born before graduation nor to potential fertility as proxied by marital status.

The remainder of this paper is organized as follows: Section 2.2 reviews the relevant literature. Section 2.3 describes the data and Section 2.4 shows the empirical approach. The results of the empirical analysis, as well as robustness checks, are presented in Section 2.5. Section 2.6 concludes the findings of the paper.

2.2 Literature review

In the following, we review what influences university graduates' decisions after the completion of their degree. This helps to select crucial control variables and helps to put our results in context. The literature dealt extensively with this question but has come up with mixed results.

Oechsle, Knauf, Maschetzke and Rosowski (2009) stress the importance of private life planning on career orientation. With increasing age, as partnerships become more important for young adults, private life and family planning have a more substantial and more significant influence on career and occupational choices. Analyzing data on American bachelor graduates, Joy (2000) finds that men hold jobs that have a higher self-reported career potential than women. Further, females are much more likely to enter white-collar work than any other occupation, regardless of their college major (Joy, 2000). Both results suggest that women might enter occupations in STEM at a lower rate than males after university. A general exit behavior is already found by Preston (1994), Preston (2004), L. A. Morgan (2000), Kahn and Ginther (2015), and Hunt (2016), while only Hunt (2016) uses both genders as well as non-STEM and STEM workers. Thereby, only she can deny a pure gender effect for the U.S. while not answering if the effect is driven by not-entering or leaving the STEM workforce.

Theoretical models, as well as empirical studies, suggest that childbearing can be a possible explanation for women ending in occupations unrelated to their degree. The human capital model of Polachek (1981) predicts that individuals who already know that they will interrupt their career and temporarily exit the labor force will choose occupations that have relatively lower skill depreciation rates. Thereby, women avoid potentially high losses of income and costly re-entries. He shows that the different amounts of time spent in the labor force can indeed explain significant parts of the differences in professional employment between men and women. Perna (2004) identifies gender as an essential factor that represents an individual's preferences in the process of occupational choice since women plan their careers in conjunction with their plans for raising children. The time during which they cannot be an active part of the labor force represents interruptions and career benefits delays. According to K. Jansen and Pascher (2013), female students anticipate the potential future problems concerning the reconciliation of work and family life and are thus underrepresented in degree-related employment. Looking at STEM students, in particular, Ivanova and Stein (2013) find that work-family balances might be one of the reasons why more women drop out of academic research in chemistry. This implies that not only having children but also the wish to have children in the near future plays an important role.

Ceci and Williams (2010) as well as Wang, Eccles and Kenny (2013) argue that women have more often both a high mathematical and a high verbal ability leading to a greater range of both STEM and non-STEM career opportunities for them to choose. Friedman-Sokuler and Justman (2016) confirm that differences in mathematical abilities cannot explain the gender gap in STEM fields. Instead, they highlight the role of cultural and psychological factors as well as social and economic incentives. Similar results are presented by Lubinski and Benbow (2006), who find that women with high math ability are less interested in pursuing a career in a math-intensive field than their male peers.

There is strong evidence that another factor influencing the choice of a career in STEM is self-efficacy, that is, how much an individual believes in his or her abilities to achieve goals or overcome obstacles (Enman & Lupart, 2000). Among others, Hübner et al. (2017) and Heilbronner (2013) have shown that there are gender differences concerning self-efficacy within STEM fields, with women reporting lower levels than men. Hence, it can be expected that women, when faced with the challenges of a STEM undergraduate degree, may not believe in their abilities to succeed in a STEM environment and discontinue their pathway in the STEM field (Heilbronner, 2013). Then even after graduation, they may question their readiness for a STEM occupation. Arcidiacono (2004), Zafar (2013), Wiswall and Zafar (2015), and Biewen and Schwerter (2019) suggest that not factors like expectations or perceived abilities but preferences explain gender differences in the college major choices.

The variety of factors above raises the presumption that women are especially likely to pursue a non-STEM occupation career while graduating in STEM. The listed results are, however, not always specific for university graduates but more general. Thus, this chapter contributes to the question of whether women, even if self-selected into STEM studies with finished degrees, still do not enter the labor market with a STEM occupation.

2.3 Data

We exploit the Graduate Panel of the German Centre for Higher Education Research and Science Studies (DZHW). The survey aims to understand better the career paths of German higher education graduates asking a variety of questions on the course of study, transition to a professional career, further education as well as sociodemographic characteristics. The survey population consists of all higher education graduates who completed a degree at a German institution of higher education in either the winter or summer semester of the respective year. Due to the unique sample and survey design, the DZHW Graduate Panel offers the best opportunities to comprehensively examine research questions about German university graduates (Baillet, Franken & Weber, 2017, 2019).

In the analysis, we pool together the observations of the 2005 and the 2009 cohort.[1] Both cohorts include graduates of traditional degree courses as well as bachelor graduates. We include information from the first and second survey waves, which are respectively one year and five years after graduation. We refer to this period as the career start of university graduates. The inclusion of the second wave helps overcome the problem of individuals taking a gap year after graduation, or individuals who delay their career start for other reasons, such as childbirth. Thus, only if the career start needs more than five years, we miss the respective graduate. Only a very small fraction of our sample was without any job in general and therefore omitted for the analysis.

Initially, the combined dataset contains 22,282 observations, of which 11,788 are from the 2005 cohort, and 10,494 are from the 2009 cohort. We restrict our sample to graduates who are older than 20 and younger than 40 at the time they finish their degree. Observations with more than one degree from different fields of study, as well as individuals who report having more than one job at the same time, are also excluded from the sample to prevent having unclear information in both explanatory and explained variables. Further, we only include those individuals in our analysis that responded to both survey waves.

The final sample used for the empirical analysis contains observations of 13,181 individuals observed at two waves. More than half of the individuals (58.92%) in the sample are female, and 13.36% are women with a STEM degree. An overview of the distribution of the nine subject groups can be found in the appendix, figure A.1. In the analysis, we look at first at STEM in general. Then, we also break this heterogeneous group into

[1] The DZHW surveys graduates only every four years and the data for 2013 is not yet publicly available.

engineering and computer sciences (EngComp) and mathematics and natural sciences (MatNat). Thereby, we follow the literature by distinguishing those two groups (for example, Hunt, 2016).

A detailed overview of the summary statistics of all covariates is shown in Table 2.1. Since the observations are pooled over both survey waves, the number of observations doubles for variables that do not contain any missing values. Most of the covariates are dummy variables that take on the value one if the statement is true and zero otherwise.

Table 2.1 – Summary statistics of covariates

	Mean	SD	Min	Max
Outcome				
Unrelated	0.1535	0.3605	0	1
Main variables of interest				
Female	0.5890	0.4920	0	1
STEM	0.3504	0.4771	0	1
MatNat	0.1218	0.3271	0	1
EngComp	0.2286	0.4199	0	1
At least one child born before graduation	0.0445	0.2061	0	1
Experiences before graduation				
Vocational training before university	0.2679	0.4429	0	1
Employment before university	0.3172	0.4654	0	1
Voluntary internship	0.3879	0.4873	0	1
Mandatory internship	0.5349	0.4988	0	1
Student assistant	0.3662	0.4818	0	1
Working student	0.3322	0.4710	0	1
Parental background				
At least one parent with Abitur	0.2262	0.4184	0	1
At least one parent with a university degree	0.2741	0.4461	0	1
At least one parent with a blue-collar occupation	0.0605	0.2385	0	1
Personal				
Age at degree completion	26.209	2.7941	21	40
Birthyear	1979.864	3.5809	1964	1988
Cohort	0.3897	0.4877	0	1
Wave	0.5085	0.4999	0	1

HEEQ in East-Germany	0.2914	0.4544	0	1
University degree in East-Germany	0.2999	0.4582	0	1
Current occupation in East-Germany	0.2490	0.4324	0	1
Foreign	0.0316	0.1748	0	1
Educational background				
Grade of HEEQ	2.2127	0.6143	0.8	4
Year of HEEQ	1999.6768	3.3588	1983	2006
Field-specific HEEQ	0.0001	0.0246	0	1
HEEQ from vocational school	0.0203	0.1411	0	1
Foreign HEEQ	0.1073	0.3095	0	1
High School at vocational school	0.0459	0.2093	0	1
Federal state of higher education entrancy qualification (HEEQ)				
HEEQ in Schleswig-Holstein	0.0316	0.1750	0	1
HEEQ in Hamburg	0.0192	0.1372	0	1
HEEQ in Lower-Saxony	0.1027	0.3036	0	1
HEEQ in Bremen	0.0081	0.0897	0	1
HEEQ in North Rhine-Westphalia	0.1645	0.3707	0	1
HEEQ in Hesse	0.0589	0.2355	0	1
HEEQ in Rhineland-Palatina	0.0369	0.1885	0	1
HEEQ in Baden-Württemberg	0.1311	0.3375	0	1
HEEQ in Bavaria	0.1492	0.3563	0	1
HEEQ in Saarland	0.0064	0.0796	0	1
HEEQ in Berlin	0.0372	0.1892	0	1
HEEQ in Brandenburg	0.0374	0.1898	0	1
HEEQ in Saxony	0.0259	0.1588	0	1
HEEQ in Mecklenburg-Vorpommern	0.0975	0.2966	0	1
HEEQ in Saxon-Anhalt	0.0391	0.1938	0	1
HEEQ in Thüringen	0.0544	0.2268	0	1
University Information				
Type of degree: Diplom	0.5830	0.4931	0	1
Grade of University degree	1.8233	0.5413	1	4
Type of degree: Magister	0.0667	0.2495	0	1
Type of degree: Bachelor	0.1646	0.3708	0	1
Type of degree: State Examination	0.0855	0.2796	0	1
Type of degree: Teaching degree	0.0972	0.2962	0	1
Type of degree: Other	0.0031	0.0557	0	1

Applied University	0.3229	0.4676	0	1
Family Information				
Partner without employment	0.1190	0.3238	0	1
In a relationship	0.4879	0.4999	0	1
Married	0.2558	0.4363	0	1

Note: The table shows summary statistics for the outcome *unrelated* and all regressors used in the empirical analysis, pooled over both survey waves. In the regressions, we further use state specific effects for the state in which the higher education entrance qualification (HEEQ) was obtained. There are 13,181 observations included in the sample.

2.3.1 Dependent variable

The construction of the dependent variable is a critical and central issue in the analysis. Using a sample of all university graduates, simply grouping the dependent variable into STEM and non-STEM occupations, is not an option, since graduates from non-STEM fields are less likely to work in STEM occupations after graduation than the other way around. Therefore, the analysis has to focus on job-relatedness. The definition of a degree-occupation match is the only way to incorporate the career paths of both STEM and non-STEM graduates into the model and to provide a comparison between these groups. However, clear definitions of degree-relatedness do not exist. With the available data, two possibilities are feasible to construct a measure of job-relatedness: one option is to compare the respective degree fields with the current occupation and determine whether an occupation is related to the field of study or not. However, an inquiry at the Federal Employment Agency revealed that there are no official matches of fields of study and occupations (apart from STEM).

Consequently, we rely on a second alternative and utilize a question that is included in both waves of the survey, asking individuals to rate how closely their field of study is related to their current occupation. The official wording of the survey question is: “Would you say that your higher education qualification matches your job, concerning the academic qualification (field of study). Rate on a scale from 1 *Yes, definitely* to 5 *No, not at all.*” Although this measure is subjective, the literature on horizontal job mismatch considers it to be sufficiently powerful (for example, Fehse & Kerst, 2007). This approach is not only in line with what most studies in the overeducation literature do but has also been successfully implemented in studies on both horizontal mismatch (Robst, 2007; Verhaest, Sellami & van der Velden, 2017) and persistence of STEM graduates in STEM occupations (Hunt, 2016; Y. J. Xu, 2013).

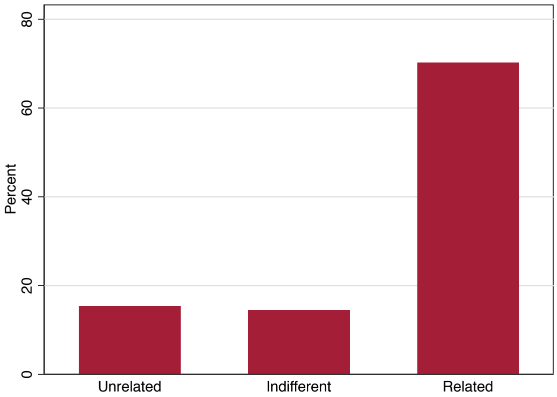
Since the intermediate points on the scale are not labeled and to simplify the interpretation a degree-related job is defined as such if individuals rated the match between a field of study and current occupation with 1 or 2. Ratings of 3 constitute the middle category of having a neither related nor unrelated (indifferent) job, and ratings of 4 or 5 are taken to define that an individual is currently holding a job unrelated to the degree (unrelated job).

Figure 2.1 gives an overview of the sample distribution of job adequacy. This view gives the first insight into the transition behavior of individuals in the sample. The graph shows that seven out of ten (70.23%) individuals in the sample reported that their current occupation is closely related to their field of study. Only 15.35% rate their job unrelated to the field of study in which they majored.[2] Almost as many, 14.42%, a similar number of participants, can be considered to be somewhat indifferent or uncertain, stating that their job is adequate to their degree field. Thus, in general, graduates are more likely to have related than unrelated occupations. Figures A1 and A2 in the appendix further split Figure 2.1 by gender and STEM. There, one can see that females, as well as non-STEM graduates, are more likely to have a degree-unrelated occupation.

Table 2.2 provides a more detailed description of the levels of job unrelatedness by STEM and non-STEM fields of study and gender. The distribution of the job unrelatedness categories among the different fields of study for all working individuals shows that STEM graduates report more often having a job that matches the field of study than non-STEM graduates (74.35% and 67.99% of individuals, respectively). Accordingly,

[2] Official numbers to assess the reliability of our measure are difficult to find. The OECD reports a field-of-study mismatch for Germany of 20%; however, data only exists for the years 2015 and 2016 (OECD, 2017).

Figure 2.1 – Reported levels of job unrelatedness in general



Note: The graph shows the reported levels of job unrelatedness in the sample, pooled over both waves. Job unrelatedness is proxied with the survey question on how closely the field of study matches the current job. For the regression analysis, we construct a binary variable in which we include the group of *indifferent* in *related*. A sensitive check including the middle category in *unrelated* confirms the results but with higher coefficients. For 2016, official OECD-data find a job-unrelatedness for Germany of around 20%, which is similar to the data in use.

having an unrelated occupation is much more of a problem for graduates of non-STEM fields (17.40%) than for STEM graduates (11.54%). Given that education in STEM fields of study is very often targeted at a specific occupation (for instance, a degree in mechanical engineering aims at preparing for a career as a mechanical engineer), these results seem plausible. Looking at the two STEM subgroups engineering and computer sciences (EngComp), and mathematics and natural sciences (MatNat), does not reveal surprises either: the reported shares for related and unrelated occupations are all roughly the same as the values for all STEM graduates. EngComp exhibits the lower shares of individuals with an unrelated occupation (10.79%), compared to MatNat (12.95%).

Examining the difference between the genders is even more insightful. Among STEM graduates, women are almost four percentage points more likely to report having a job that is not related to their degree field. Although female graduates of non-STEM fields are also more likely to report an unrelated occupation, the gender difference is much smaller in this group. Again, this notion also translates to the opposite category of having a closely related job to one's university major. Among both groups, STEM and non-STEM graduates, women are also less likely to have a related job, with the difference again being more substantial for STEM than for non-STEM graduates. When looking at the STEM subgroups, it becomes once more apparent that a more differentiated perspective provides essential information. Within the EngComp graduates group, women are more likely to have a job that is unrelated to their STEM degree and less likely to have a job that is entirely related. In MatNat, the result is similar, while the difference is much smaller. There is also no gender difference in having a closely related occupation in MatNat.

Table 2.2 – Job adequacy by field of study and gender

	Male %	Female %	All %
All STEM			
Unrelated	10.15	13.78	11.54
Indifferent	14.04	14.24	14.12
Related	75.81	71.98	74.35
Math and Natural Sciences			
Unrelated	12.83	13.03	12.95
Indifferent	12.01	12.73	12.45
Related	75.16	74.25	74.60
Engineering and Computer Sciences			
Unrelated	9.43	14.77	10.79
Indifferent	14.59	16.21	15.00
Related	75.98	69.02	74.21
All other fields			
Unrelated	17.84	17.22	17.40
Indifferent	15.07	14.42	14.61
Related	67.10	68.37	67.99

Note: The table shows the relationship between job relatedness and degree of study, pooled over both survey waves. Graduates are grouped by STEM, the subgroups Math and Natural Sciences, and Engineering and Computer Sciences, and all other fields for comparison. The percentages are presented for men, women and both together. Here, we get a first impression of the differences between the subjects and the gender. The corresponding survey question was: “Would you say that your higher education qualification matches your job, with respect to the academic qualification (field of study). Rate on a scale from 1 *Yes, definitely* to 5 *No, not at all.*” Ratings of 1 and 2 have been converted to *related job*, a rating of 3 to *indifferent*, and ratings of 4 and 5 to *unrelated job*.

For the empirical analysis that will follow in Section 2.5, the information of the three-categorical degree relatedness measure is further condensed into a binary variable, *unrelated*. The outcome variable is equal to one if the current job is not adequate (or unrelated) to the field of study and equal to zero otherwise, that is, if a job is somewhat or entirely adequate (or related). Table 2.2 shows that the middle category does not differ much across either field of study or gender. Because of that we dichotomizing the variable (ratings of 5 and 4 as an unrelated job, $y_{it} = 1$, and ratings of 1 - 3 as a related job, $y_{it} = 0$). This simplification eases the interpretation of the results. To show the robustness of the results, we will further show an estimation in which (i) the middle category is coded as an unrelated occupation, and (ii) the 5-likert scale variable is not simplified.

2.4 Econometric model

We aim to examine whether females with a STEM degree are less likely to enter the job market with a STEM occupation or not and whether this is most pronounced for females with children. Thus, we will examine whether female STEM graduates are not entering STEM occupations at a higher rate than men, relative to other professional fields. The baseline regression equation is as follows:

$$y_{it} = \alpha_1 + \gamma_1\, female_i + \delta_1\, STEM_i + \rho_1\, STEM_i\, female_i + \boldsymbol{\beta}_1'\, \mathbf{X}_{it} + \epsilon_{it} \tag{2.1}$$

where index i stands for the individual and t for the wave. The dependent variable y_{it} is a binary variable that is equal to one if an individual i holds a job unrelated to their degree in time t and equal to zero if the job is related.

On the right-hand side of Equation (2.1), *female* is a dummy variable equal to one if the observed individual is a woman and equal to zero if the individual is a man. The binary variable *STEM* is equal to one if an individual has a degree in STEM and zero for individuals holding degrees from all other fields of study. The classification of STEM degree subjects follows the classification of the Federal Employment Agency (Bundesagentur für Arbeit, 2019).[3] We include the interaction of the two dummy variables in the regression equation to isolate the specific effect for female STEM graduates. The coefficient on ρ_1 hence gives the field-specific gender difference. Adding δ_1 and ρ_1 shows the changed probability of working in an unrelated job specifically for female STEM graduates. A positive value of this sum would indicate that female STEM graduates not entering STEM occupations excessively. In contrast, the sum's negative value would indicate a lowered entry rate of females compared to males within STEM.

The vector of variables $\mathbf{X}_{it}$ includes several covariates to control for other factors that might affect the attrition from the STEM field. The covariates are derived from both the literature on job mismatch and STEM entry behavior. Table 2.1 lists the full set of covariates. We include demographic information, (work) experience before graduation, socio-cultural variables, personal and educational background, as well as job characteristics, study information, and origin information. Additionally, following Hunt (2016), we include dummies for all other areas of studies.[4]

We further follow Hunt (2016) by separating STEM into the groups engineering and computer sciences (EngComp) and mathematics and natural sciences (MatNat). For this, we replace the STEM dummies with a dummy for EngComp and for MatNat. The regression equation looks as follows:

$$\begin{aligned} y_{it} = \alpha_2 + \gamma_2\, female_i + \delta_2\, EngComp_i + \rho_2\, (EngComp_i \cdot female_i) \\ + \mu_1\, MatNat_i + \mu_2\, (MatNat_i \cdot female_i) + \boldsymbol{\beta}_2'\, \mathbf{X}_{it} + \eta_{it} \end{aligned} \tag{2.2}$$

To investigate possible fertility effects, we further add a children-dummy into the equation and interact the dummy with the variables for the EngComp degree and gender

[3] The Federal Employment Agency includes the fields mathematics, physics, chemistry, pharmacy, biology, geo-sciences geography, as well as, computer sciences, and all engineering fields in STEM.

[4] Results are robust to not adding all other areas of study.

and their interaction:

$$\begin{aligned} y_{it} = \alpha_3 &+ \gamma_3 \, female_i + \delta_3 \, EngComp_i + \rho_3 \, (EngComp_i \cdot female_i) \\ &+ \lambda_3 \, (EngComp_i \cdot female_i \cdot children) + \gamma_4 \, children \\ &+ \gamma_5 \, (children_i \cdot female_i) + \gamma_6 \, (EngComp_i \cdot children_i) \\ &+ \mu_3 \, MatNat_i + \mu_4 \, (MatNat_i \cdot female_i) + \boldsymbol{\beta}_3' \, \mathbf{X}_{it} + \xi_{it} \end{aligned} \tag{2.3}$$

For the variable *children*, we use three different specifications. To evade possible endogeneity of children and occupation decisions, our main variable *children* is a binary variable equal to one if an individual has one or more children before graduation and zero otherwise. Concentrating on children born before leaving university mitigates the potential endogeneity because we know childbirth happened before obtaining the degree. Otherwise, entering the job market and the willingness to get a child might be simultaneous. Then, we further use the variable *children*^{W} which is additionally equal to one in the second wave when the graduate got children before that wave. Lastly, to check for robustness, we further neglect the timing of birth and just set the variable equal to one if the graduate has one or more children in the respective wave. This, however, has to be interpreted with great caution.

The interaction of *EngComp* and *female* still isolates the field-specific gender difference that is specific for female EngComp graduates, but now only for childless women. The coefficient of the triple interaction, λ_3, shows if the field-specific gender difference effect for women with children. Thus, if $\delta_3 + \rho_3 + \lambda_3 > \delta_3 + \rho_3 > 0$, women with an EngComp degree would over-proportionately not enter their occupational field because of childcare obligations, relative to men, relative to other professional fields and relative to those without children.

The job-entry and wish to have children might also be affected by the relationship-status of individuals. Employees might fear that young women will have children in the near future. Moreover, women might already be planning to have children soon after graduation and, thus, do not start an occupation in the field of the degree but *an unrelated* job. This might be more likely for married women, which is why we also use a marriage-dummy instead of the children-dummy (and the respective interactions) to check for potential-fertility effects, similar to Biewen and Seifert (2018) and S. O. Becker, Fernandes and Weichselbaumer (2019).

We estimate the regression equation using a pooled OLS regression model.[5] Even though the outcome variable is binary, the interpretation of (triple) interaction terms of a non-linear logit model are non-trivial. Standard errors are clustered at the individual level and account for heteroskedasticity.

2.5 Empirical results

The following section presents the results of the pooled OLS regression analysis. The outcome variable *unrelated* is equal to zero if the occupation is related and equal to one if the occupation is unrelated. A negative effect, thereby, shows that individuals are more likely to have a degree-related occupation.

2.5.1 Baseline results

Table 2.3 shows the basic regression results of Equations (2.1) to (2.3). The first column reports estimation results of Equation (2.1) without control variables; column (2) with controls; column (3) reports results distinguishing between EngComp and MatNat, and column (4) includes the children variable and its respective interactions. Then, column (5) includes being married to account for the potential wish to have children soon, and column (6) includes both children and being married.

First, in column (1), we see a negative STEM effect of -7.7% at the 0.1% significance

[5] Estimations using random-effects or logit models lead only to minor changes after the third decimal and do not influence the interpretation of the coefficient nor the significance level.

Table 2.3 – Baseline regression results

	Dependent variable: Degree unrelated occupation					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	−0.006	0.011	0.011	0.011	0.005	0.006
	(0.010)	(0.010)	(0.010)	(0.011)	(0.012)	(0.012)
STEM	−0.077***	−0.064***				
	(0.011)	(0.014)				
STEM×Female	0.043**	0.050**				
	(0.015)	(0.015)				
EngComp			−0.078***	−0.079***	−0.088***	−0.088***
			(0.014)	(0.014)	(0.015)	(0.015)
EngComp×Female			0.052**	0.055**	0.048*	0.050*
			(0.018)	(0.019)	(0.020)	(0.020)
MatNat			−0.007	−0.007	−0.007	−0.007
			(0.019)	(0.019)	(0.019)	(0.019)
MatNat×Female			0.011	0.011	0.010	0.011
			(0.021)	(0.021)	(0.022)	(0.022)
Children				−0.022		−0.018
				(0.040)		(0.040)
Female×Children				0.006		−0.002
				(0.047)		(0.047)
EngComp×Children				0.025		0.012
				(0.054)		(0.054)
EngComp×Female×Children				−0.074		−0.069
				(0.086)		(0.087)
Married					−0.016	−0.015
					(0.014)	(0.014)
Female×Married					0.022	0.023
					(0.017)	(0.017)
EngComp×Married					0.036^{+}	0.036^{+}
					(0.020)	(0.020)
EngComp×Female×Married					0.016	0.019
					(0.042)	(0.042)
Control variables	No	Yes	Yes	Yes	Yes	Yes
Female Degree Effect	−0.034	−0.014	−0.026	−0.025	−0.039	−0.038
Adjusted R^2	0.007	0.081	0.082	0.082	0.082	0.082
Observations	13177	13177	13177	13177	13125	13125

Note: Pooled OLS with individual clustered and heteroskedastic robust standard errors in parentheses. Control variables includes all variables listed in Table 1. 'Female Degree Effect' is the sum of 'STEM' and 'STEM×female' for column (1) and (2) and for 'EngComp' and 'EngComp×female' in column (3) to (6). 'Children' is short for 'at least one child born before graduation'. Regression results showing all coefficients is available upon request. $^{+}$ (p < 0.10), * (p<0.05), ** (p<0.01), *** (p<0.001) level on unrelatedness. Higher relatedness probability is reduced for females because the interaction term has a positive coefficient of 4.3 percentage points, significant at the 1% level. Thus, male STEM-graduates are less likely to have an unrelated job. Adding both named effects gives the female STEM effect equal to -3.4%, shown at the bottom of the regression tables. Thus, female STEM graduates, as well as males, are more likely to have a degree-related occupation, but it is lowered for women by 4.4 percentage points, from -7.7 to -3.3%. If we include the control variables in column (2), which are summarized in Table 2.1, the STEM coefficient increases slightly to -6.3% and the field-specific gender difference coefficient increases to 5.0 percentage points. These changes leave a female STEM effect of only -1.3% compared to men and women of all other fields. Men, thereby, have a higher probability of working in a degree-related occupation if they have a STEM degree, compared to other men. Women experience this as well, but to a much lesser extend, if at all.

Next, we want to see which subgroups are the main driver of the STEM effect. Column (3) shows that the EngComp-graduates drive the effects. The coefficients for MatNat and MatNat×female are close to zero. Column (4) then investigates whether this EngComp-specific gender differences can be explained by fertility. As shown in the literature review, this question is so far answered ambiguously in general and almost not at all for the career start. Including the variable *children* (at least one child is born before graduation) and its interactions, we see that the general effects are robust, and we cannot detect any children-specific effects (see column 4).[6] This would suggest that having children before graduation is not essential for the job-entry decision. We also include being married in columns (5) and (6) to control for potential fertility in the near future. While the variable could be endogenous, since we do not observe when individuals got married, it does not alter the estimation results much.

In the following, we will stick to the separation of STEM into EngComp and MatNat, because the main driver of the STEM effect is EngComp.

[6] If we do not include control variables, there is a significant child-specific effect. This effect is, however, explained by the control variables.

2.5.2 Sensitivity & robustness

In Table 2.4, we check if the basic regression results are sensitive to the recoding of the outcome variable (column 1 and 2), subsamples of wave 1 (column 3) and wave 2 (column 4), and the current career outlook (column 5). Column (6) and (7) further include two different specifications of the children variable.

When recoding the middle category for the 5 points Likert scale "indifferent" to count as an unrelated occupation as well, the coefficients increase slightly. Further, the triple interaction becomes marginally significant at the 10% level with a counterintuitive negative coefficient. Women with a degree in EngComp and children born before their mothers' graduation seem to have an advantage compared to women without children born before graduation. This suggests that women who select themselves into EngComp and already have children before graduation might be more deterministic and stay within their field. However, the triple interaction is not significant anymore once we look at the 5 points Likert scale variable in column (2). Thus, the significance in column (1) is only weak evidence. The male degree estimate, as well as the field-specific gender difference, are robust.

Next, for the current occupation in wave 1 in column (4), we see a pronounced male effect for a degree in EngComp by 11.1% at the 0.1% significance level and a field-specific gender difference, which is higher compared to the regressions before, namely 6.2 percentage points. When only the male effect shrinks towards zero and the field-specific gender difference is constant, the actual female effect to have an unrelated occupation decreases as well.

For wave two (column 4), however, the estimates are lower. Men have a lowered but still negative probability of having an unrelated job by -4.5 percentage points, the lowest so far. Further, the EngComp-specific gender difference is almost equal in absolute value.

Next, we only keep individuals who report their current job either as a mid-term or

Table 2.4 – Sensitivity check

	Dependent variable: Degree unrelated occupation						
	Changed middle category	5-likert scale	Wave 1	Wave 2	Outlook	Children before specific wave	Children without time information
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.017	0.012	0.004	0.018	0.009	0.012	0.005
	(0.0130)	(0.0359)	(0.0146)	(0.0122)	(0.0106)	(0.011)	(0.011)
EngComp	−0.100***	−0.318***	−0.114***	−0.045**	−0.063***	−0.079***	−0.083***
	(0.0188)	(0.0493)	(0.0201)	(0.0166)	(0.0146)	(0.014)	(0.015)
EngComp×Female	0.069**	0.201**	0.066*	0.044*	0.050*	0.055**	0.062**
	(0.0243)	(0.0645)	(0.0264)	(0.0222)	(0.0195)	(0.019)	(0.020)
MatNat	−0.031	−0.095	−0.051^{+}	0.037	0.017	−0.007	−0.008
	(0.0251)	(0.0680)	(0.0269)	(0.0249)	(0.0212)	(0.019)	(0.019)
MatNat×Female	0.010	0.055	0.025	−0.003	−0.007	0.011	0.011
	(0.0275)	(0.0771)	(0.0286)	(0.0275)	(0.0232)	(0.021)	(0.022)
Children	−0.024	−0.164	−0.061	0.013	−0.015		
	(0.0440)	(0.1188)	(0.0524)	(0.0485)	(0.0412)		
Female×Children	0.023	0.137	0.042	−0.027	−0.007		
	(0.0518)	(0.1431)	(0.0613)	(0.0568)	(0.0483)		
EngComp×Children	0.059	0.207	0.086	−0.031	0.006		
	(0.0687)	(0.1772)	(0.4421)	(0.0642)	(0.0514)		
EngComp×Female×Children	−0.183^{+}	−0.419	−0.092	−0.059	−0.020		
	(0.0958)	(0.2675)	(0.1427)	(0.0980)	(0.0957)		
Children^{W}						−0.015	
						(0.039)	
Female×Children^{W}						0.000	
						(0.045)	
EngComp×Children^{W}						0.025	
						(0.052)	
EngComp×Female×Children^{W}						−0.077	
						(0.083)	
Children^{E}							−0.003
							(0.016)
Female×Children^{E}							0.028
							(0.019)
EngComp×Children^{E}							0.022
							(0.022)
EngComp×Female×Children^{E}							−0.052
							(0.041)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subsample	-	-	1st Wave	2nd Wave	Outlook	-	-
Female EngComp Effect	−0.031	−0.118	−0.048	−0.001	−0.013	−0.024	−0.021
Adjusted R^2	0.091	0.127	0.081	0.075	0.071	0.082	0.083
Observations	13177	13177	6477	6700	10808	13177	13177

Note: 'Changed middle category' refers to the decision of the recoding of the 5 points likert scale variable. In column (1), apart from all other regressions, we include the middle category to unrelated. Column (2) leaves out the recoding of the outcome variables and just uses the 5 point likert scale. Wave 1 and Wave 2 use only the information of the graduates from the first or second survey wave. The outlook subsample only includes individuals who see the current job as mid- or longterm outlook, that is, we dropped the short-term. Children^{W} refers the dummy equal to one for wave one if the child was born before graduation, and for wave two if the child was born ins ave one. Children^{E} neglects the timing of birth. Thus, the results for Children^{E} should be interpreted with cautious because of likely endogeneity. *Female EngComp Effect* is the sum of 'EngComp' and 'EngComp×female'. Individual clustered and heteroskedastic robust standard errors in parentheses. $^{+}$ (p < 0.10), * (p<0.05), ** (p<0.01), *** (p<0.001) long-term solution in column (5). Here, the estimation results are again very similar to the estimates of Table 2.3. Lastly, we use two different variations of the children variable: $children^W$ refers the dummy equal to one for wave one if the child was born before graduation, and for wave two if the child was born in wave one. The variable $children^E$ neglects the timing of birth. Thus, the results for $children^E$ should be interpreted with caution because of likely endogeneity. The regression results are, again, fairly robust.

Then, we check whether the estimation is robust to the inclusion of additional variables. Therefore, for each column separately in Table A1, we add the following variables: if the graduate has a partner who is employed, if the graduate is in a relationship, if the graduate has the desire to have children,[7] the child age and the share of males in the area of the degree of study. In each regression, we further include an interaction of the named variables with the gender dummy. The estimation is very robust, and we see hardly any changes in our main variables of interest.

2.5.3 Internship and other relatedness variables

One idea to explain different labor market transitions of men and women is to see whether they had an internship first before entering regular employment. For the probability of an internship right after graduation, we use only the observations of the first wave. Here, we see in Table A2 that EngComp graduates, in general, are less likely to enter the labor market with an internship irrespective of the gender of by 5% at the 0.1% significance level. However, in general, irrespective of the field, females are more likely to have first an internship with a higher probability of 3% at the 1% significance level.

Next, we analyzed whether the current job position (columns 3 and 4) and job task level (columns 5 and 6) are related to the degree. For both, we find a general advantage for EngComp graduates and a general disadvantage for females. The field-specific gender

[7] This information comes from the second wave of the survey, and we assume that the desire is constant over both waves.

difference, though, is only significant for the job position. Thus, female EngComp graduates are more likely to have a job position lower to what their degree should have made available. Women with a degree in EngComp are thus less likely to have a degree-related job and have to start lower on the career ladder. In terms of the level of job tasks, female EngComp students are not worse off compared to their male counterparts, only females in general.

2.5.4 Discussion of the results

The regression analysis has shown that there is a higher rate of non-transitions to EngComp (and thereby STEM) occupations for women compared to men. The EngComp work environment, therefore, seems to be significantly less attractive to its female graduates compared to males. The second part of the empirical analysis has shown that there do not seem to be non-entries due to child care responsibilities.

The literature review showed that not mathematical abilities (Friedman-Sokuler & Justman, 2016) but differences in self-efficacy (Heilbronner, 2013), personal interest (Heilbronner, 2013), and willingness to compete (Buser, Peter & Wolter, 2017) are reasons for the gender difference in STEM and EngComp. However, the graduates in our sample already self-selected themselves into STEM or EngComp studies, showing some confidence and interest in those fields. One possible explanation of the results might be that the lower, for example, self-efficacy is still prevalent.

Besides the internal factors mentioned previously, external factors such as experiences at the workplace and the availability of role models also play an essential role in the decision to pursue a career in the STEM field (Heilbronner, 2013). Studies unambiguously conclude that women still face substantial barriers and discrimination at STEM workplaces (Danbold & Huo, 2017). The prevailing perception in technological occupations is the belief that being family-oriented is not associated with professional success. This assumption is often described as a *family penalty*, where women who wish to do both, climb the career ladder and raise their children, often experience their family responsibilities as a barrier to advancement (Simard et al., 2013). It seems that for women to achieve career goals in technological professions, they have to delay getting children or refrain from having children at all (Simard et al., 2013). Additionally, Danbold and Huo (2017) show that men undermine the success of women in STEM. The external factors, however, will not be testable with our data.

Overall, the gender gap in the EngComp field predicted by the model does exist but is small. This gap might discourage female EngComp graduates from entering EngComp occupations in the future, creating a vicious circle in the attraction of female talent to the STEM sector due to missing role models.

The missing fertility effect is in line with, for example, Polachek (1981) and K. Jansen and Pascher (2013). Females prioritizing a balanced work and family life might not even start to study EngComp in the first place.

One should note the pre-selection step into the fields of study, which is why we should see the coefficients as lower-bounds. The missing effect of fertility measured by us is not necessarily causal. However, if the variable *children* were indeed endogenous, we would expect it to be positively correlated with degree-non-relatedness. This might be interpreted as an indication that there is indeed no causal effect of fertility, as our measured effect would be a lower bound. Still, the true effect of fertility would not be negative.

2.6 Conclusion

We use a sample of German university graduates from all fields of study to determine whether there is a specific non-transition for women from STEM fields of study. Therefore, we focus on the job-entry decision only and not general exit studies as done by Hunt (2016). Non-transitions are considered as such when a graduate reports holding a job unrelated to his or her field of study within the first five years after graduation.

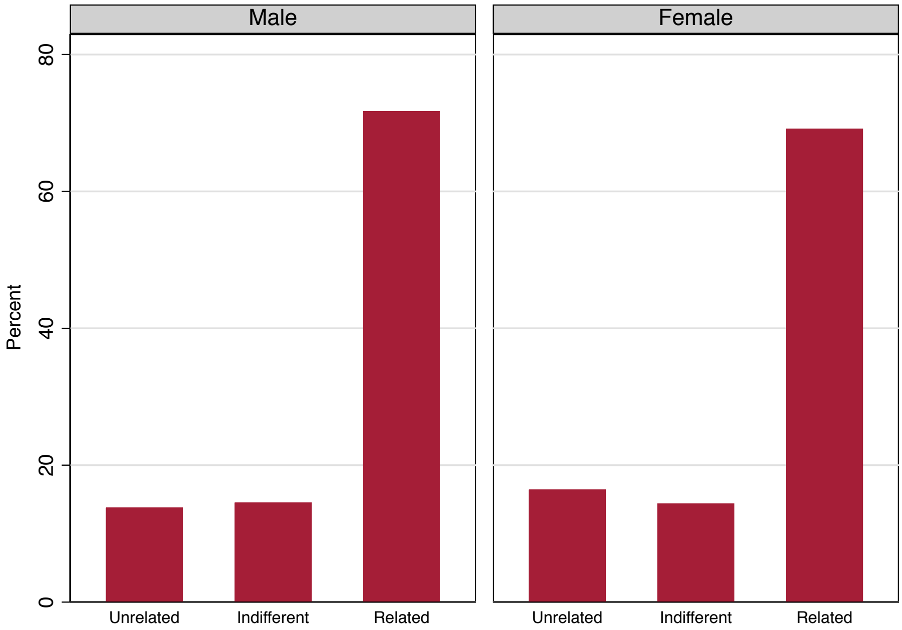
The regression analysis shows that graduates from EngComp (and thereby STEM) are more likely to enter an occupation related to their study compared to other fields. Second, this higher transition-rate is reduced for women for EngComp occupations compared to men and relative to women of other professional fields. This result is robust to different model specifications and more pronounced in the first wave. For the second wave, we even find no STEM advantage for women anymore.

Neither children nor potential fertility, as proxied by marriage, do help to explain the degree-unrelatedness of men or women. Other reasons, such as statistical discrimination against women in EngComp, could not be tested with the usage of the data and should be focused on in future research.

Appendix

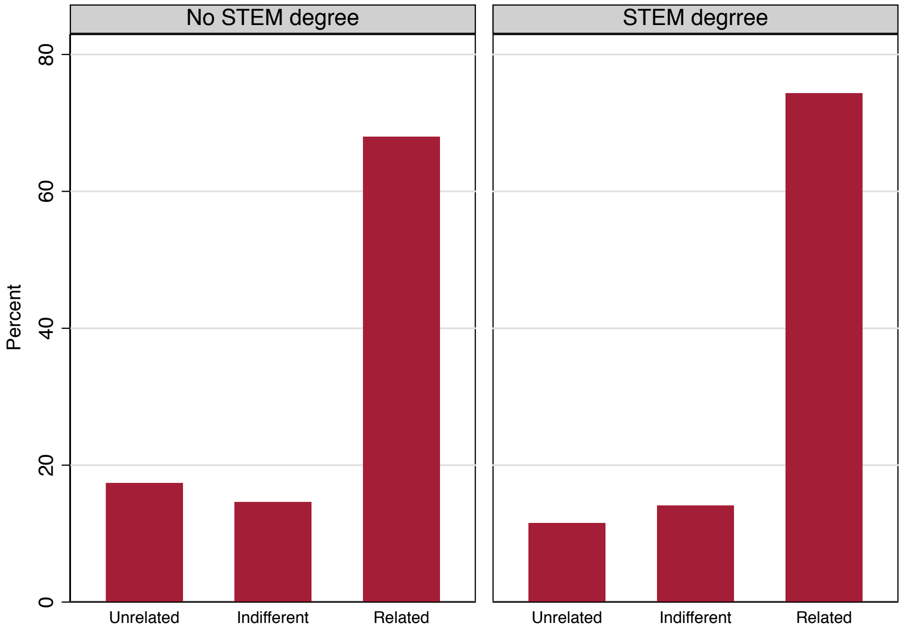
A.1 Figures

Figure A1 – Reported levels of job unrelatedness between gender



Note: The graph shows Figure 2.1 conditional on gender, i.e. the distribution of the relatedness for males on the left and for females on the right. The shares are similar, but there are slightly more males with a degree-related occupation compared to females. The middle category, *Indifferent* is the same for both.

Figure A2 – Reported levels of job unrelatedness between STEM and non-STEM



Note: The graph shows Figure 2.1 conditional on STEM-degree, i.e. the distribution of the relatedness for graduates without a STEM degree on the left and for graduates with a STEM degree on the right. The shares are similar, but there are slightly more graduates with a STEM degree having a degree-related occupation compared to non-STEM graduates. The middle category, *Indifferent* is the same for both.

A.2 Tables

Table A1 – Robustness check

	Dependent variable: Degree unrelated occupation				
	(1)	(2)	(3)	(4)	(5)
EngComp	0.010	0.016	0.029^{+}	0.011	0.021
	(0.0109)	(0.0119)	(0.0171)	(0.0106)	(0.0193)
EngComp×Female	−0.079***	−0.079***	−0.080***	−0.080***	−0.097***
	(0.0143)	(0.0143)	(0.0146)	(0.0143)	(0.0193)
Female	0.055**	0.053**	0.047*	0.055**	0.069**
	(0.0187)	(0.0187)	(0.0189)	(0.0187)	(0.0248)
Children	−0.007	−0.007	−0.009	−0.009	−0.010
	(0.0194)	(0.0194)	(0.0198)	(0.0194)	(0.0198)
Children×Female	0.011	0.011	0.008	0.011	0.017
	(0.0215)	(0.0215)	(0.0219)	(0.0215)	(0.0221)
EngComp×Children	−0.022	−0.023	−0.024	0.006	−0.021
	(0.0399)	(0.0400)	(0.0411)	(0.0498)	(0.0399)
EngComp×Children×Female	0.006	0.002	−0.002	−0.013	0.005
	(0.0468)	(0.0469)	(0.0480)	(0.0628)	(0.0468)
MatNat	0.026	0.025	0.024	0.031	0.027
	(0.0536)	(0.0536)	(0.0542)	(0.0556)	(0.0536)
MatNat×Female	−0.073	−0.068	−0.068	−0.078	−0.076
	(0.0859)	(0.0861)	(0.0866)	(0.0863)	(0.0861)
Partner employed	−0.002				
	(0.0122)				
Partner employed×Female	0.015				
	(0.0201)				
In a relationship		−0.007			
		(0.0093)			
In a relationship×Female		−0.009			
		(0.0124)			
Desire to have children			−0.009		
			(0.0123)		
Desire to have children×Female			−0.021		
			(0.0168)		
Childage				−0.003	
				(0.0042)	
Childage×Female				0.002	
				(0.0051)	
Share of males					0.048
					(0.0367)
Share of males×Female					−0.028
					(0.0472)
Control variables	Yes	Yes	Yes	Yes	Yes
Female EngComp Effect	−0.025	−0.026	−0.033	−0.025	−0.029
Adjusted R^2	0.082	0.082	0.082	0.083	0.082
Observations	13177	13125	12716	13129	13177

Note: Pooled OLS with individual clustered and heteroskedastic robust standard errors in parentheses. Control variables includes all variables listed in Table 1. 'Female EngComp Effect' is the sum of 'EngComp' and 'EngComp×female'. 'Children' is short for 'at least one child born before graduation'. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2 – Internship and other relatedness-variables

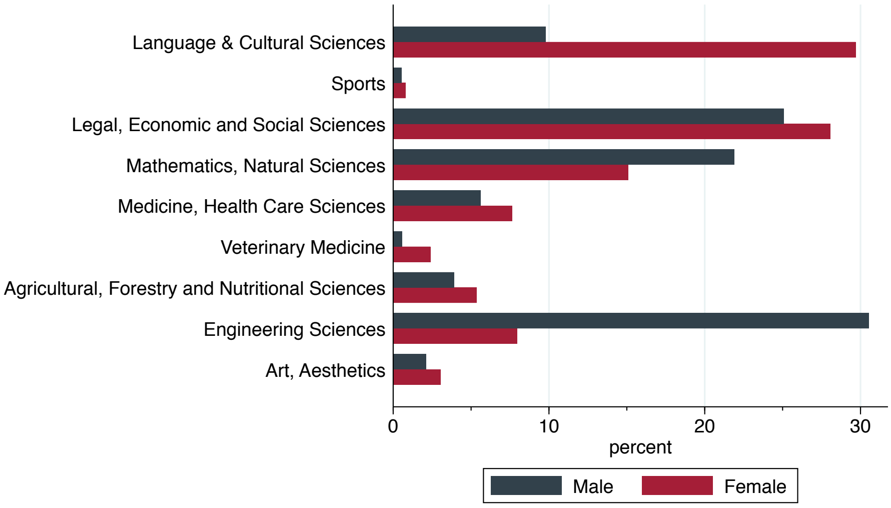
	Dependent variable:		
	Internship	Degree unrelated position	Degree unrelated task
	(1)	(2)	(3)
EngComp	-0.052^{***}	-0.086^{***}	-0.052^{***}
	(0.015)	(0.013)	(0.013)
EngComp×Female	0.018	0.053^{**}	0.023
	(0.023)	(0.018)	(0.017)
Female	0.037^{**}	0.027^{**}	0.043^{***}
	(0.012)	(0.010)	(0.010)
MatNat	−0.014	−0.021	−0.017
	(0.022)	(0.018)	(0.017)
MatNat×Female	−0.022	0.024	0.011
	(0.026)	(0.019)	(0.019)
Children	−0.027	-0.070^{*}	−0.038
	(0.033)	(0.029)	(0.035)
Children×Female	0.022	0.073^{+}	0.008
	(0.041)	(0.038)	(0.041)
EngComp×Children	0.005	0.031	0.021
	(0.040)	(0.041)	(0.051)
EngComp×Children×Female	−0.074	0.146	−0.048
	(0.080)	(0.110)	(0.087)
Control variables	Yes	Yes	Yes
Female EngComp Effect	−0.034	−0.033	−0.029
Adjusted R^2	0.085	0.095	0.065
Observations	6406	13167	13167

Note: The outcome internship is equal to one if graduates got an internship after graduation before regular employment. Job position and task level both are also subjective measures about how they related to the degree without being field-specific. 'Female EngComp Effect' is the sum of 'EngComp' and 'EngComp×female'. Individual clustered and heteroskedastic robust standard errors in parenthesis. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.3 Supplementary information

Figure A3 gives an overview of the distribution of the nine subject groups across gender and reveals that huge gender differences exist. Only around 8% of women graduate with a degree in Engineering, compared to around 30% of men. A similar trend can be seen in Mathematics and Natural Sciences. Even though the difference is not as big as in Engineering, it is still considerable: only 15% of women compared to 22% of men graduate with a degree in this field. This gives a first hint to a potential underrepresentation of women in these occupational fields. The lack of role models and the fear of discriminatory work environments might consequently discourage women from actually entering the occupational field. By contrast, figure A.1 also identifies subject groups that seem to be typically feminine, such as the field of Language and Cultural Sciences. For the analysis, we group the fields of study of math and natural sciences to MatNat, and engineering and computer sciences to EngComp.

Figure A3 – Distribution of degree subject groups across gender



Note: The graph shows the distribution of subject groups across gender. Except for the differences in Sports and in Legal, Economic and Social Sciences, all differences are statistically highly significant at the 1% level ($p < 0.01$).

Chapter 3

Does more math in high school increase the share of female STEM workers? Evidence from a curriculum reform[*]

[*] This chapter is based on: Biewen, M. and J. Schwerter (2020): Does more math in high school increase the share of female STEM workers? Evidence from a curriculum reform, unpublished manuscript, University of Tübingen.

3.1 Introduction

Recent technological changes strongly suggest that future economic growth can primarily be expected in fields related to science, technology, engineering, and mathematics (STEM) (OECD, 2010). One way to promote this growth is to foster female participation in STEM subjects with the goal of increasing the number of female STEM graduates and female STEM workers. In addition to this macro perspective, STEM-related jobs are usually well-paid due to their relation to high productivity sectors of the economy. Attracting more females into STEM subjects can, therefore, be seen as a way to improve women's career prospects and reduce the gender wage gap and gender-related earnings inequality over the life-cycle (OECD, 2007).

In this chapter, we exploit an exogenous shock in the form of a curriculum change at the high school level to investigate whether and to what extent it is possible to influence the share of female STEM graduates by extending the math and natural science content in high school. The reform was intended to improve students' general preparation for university studies and the labor market in Baden-Württemberg, one of the German federal states (Schavan, 1999). However, its largest component was an increase in math and natural science classes during the last two years of high school, which affected females more than males.

The literature cites several possible reasons for the low share of females in STEM subjects in various stages of the educational system: (i) ability (Berlingieri & Zierahn, 2014; Friedman-Sokuler & Justman, 2016), (ii) tastes and preferences (Arcidiacono, 2004; Ceci & Williams, 2010), (iii) stereotypes (Cheryan, 2012; Franceschini, Galli, Chiesi & Primi, 2014), (iv) path decisions in school (Broecke, 2013; Justman & Mendez, 2018), (v) dropping out of a STEM study programme (Ehrenberg, 2010; Kokkelenberg & Sinha, 2010), (vi) underrepresentation in university faculties (Carrell, Page & West, 2010; Griffith, 2010), and (vii) failure to transform a STEM degree into a STEM occupation in the labor market (Danbold & Huo, 2017; Sassler et al., 2017).

This chapter addresses aspect (iv) and (vii). First, by taking advantage of an exogenous shock in path decisions, we evaluate whether a reform at the end of high school changes university study decisions. We do this by evaluating the number of STEM degrees obtained by high school graduates before and after the reform using the other federal states as a control group. Second, we observe graduates' transition into the labor market. We are thus able to test whether the reform led to changes in the eventual share of females entering STEM occupations.

Our analysis is based on underused data about German university graduates collected by the German Center for Higher Education and Science Studies (DZHW). The DZHW surveys provide repeated representative samples of graduates from tertiary education institutions in Germany. We focus on the survey cohorts 2005/2006 and 2013/2014, who were surveyed one year after graduation. The surveys include background information on secondary education, tertiary education decisions as well as information on the transition to the labor market after graduation (Baillet et al., 2017, 2019). Our analysis follows the recent literature, for example, Justman and Mendez (2018), by further distinguishing between all STEM fields combined and the subfields mathematics and natural sciences (MatNat) as well as engineering and computer sciences (EngComp).

This chapter makes the following contributions. First, to the best of our knowledge, the study is one of the first to use quasi-experimental variation to study the effects of curricula changes on the inflow into STEM occupations, and one of the few to use quasi-experimental information to evaluate the effects of curricula changes on college major choices (see the detailed literature review below). Second, in contrast to many of the quasi-experimental interventions studied in the literature, the reform we are considering is unique and comprehensive in that it affected *all students* in the last two years of high school by making advanced math courses, which were chosen by only 20% of students before the reform, compulsory. This meant that a very large proportion of the population were compliers to the reform and that it 'leveled the playing field' between the genders and students of different ability levels (Domina & Saldana, 2012).

This is in contrast to other interventions considered in the literature, which often reached considerably smaller fractions of the population or only subgroups with certain ability levels (for example, De Philippis, 2017; Jia, 2016; Joensen & Nielsen, 2016). Given that our quasi-experiment affected a very large group of individuals, one of our aims is to describe potential heterogeneity of reform effects across population subgroups. Finally, our results appear to be relevant in the sense that we provide some evidence for *negative* effects of increasing exposure to math and natural sciences on later STEM participation. In this respect, we add to the literature that finds unintended consequences of interventions that were thought to equalize the opportunities of men and women (for example, Brenoe & Zölitz, 2020; Hübner et al., 2017).

The structure of this chapter is as follows. Section 3.2 discusses the related literature. In Section 3.3, we describe the institutional background and the reform in more detail. Sections 3.4 and 3.5 provide details about our econometric approach and data. In Section 3.6, we present and discuss our empirical results. Section 3.7 is the conclusion.

3.2 Related literature

A considerable literature has looked at college major choice and its determinants (for an overview, see Altonji, Blom & Meghir, 2012). Here, we only focus on articles that address the question of STEM vs. non-STEM majors.[1] In an early contribution based on controlling for observables, Levine and Zimmerman (1995) consider the effects of taking more high-school math on wages, college majors and gender-traditional occupations. They find that, for women, more math was associated with a higher likelihood of completing a technical degree and working in a technical job or a job traditional for one's sex. Arcidiacono (2004) estimates a sophisticated structural choice model of college major choices, incorporating aspects such as learning about one's abilities and uncertainty in educational outcomes. He found that math ability (but not verbal ability)

[1] The following literature review only focusses on the effects of school curricula on educational and economic outcomes. We do not review articles that specifically deal with STEM occupations in the labor market (Spitz-Oener & Priesack, 2018) or the more general topic of women in STEM (Kahn & Ginther, 2018).

is important for selecting certain majors (especially STEM), but that ability differences are far from enough to account for observed choices. Rather, differences in job and school preferences dominate college major choices. This is in line with Card and Payne (2017), who study STEM major choices in relation to an index of STEM readiness at the end of high school. They show that men and women do not differ in STEM readiness before university, but that males not interested in STEM subjects are less likely to start university studies.

The chapter connects to a wider literature analyzing the effects of differences in school curricula, especially with regard to math and the natural sciences, on later educational and economic outcomes. A number of papers have studied the effects of differential exposure to math curricula on later decisions in high school and on college attendance. For example, Aughinbaugh (2012) finds that a more rigorous high school math curriculum is associated with a higher probability of attending college. Broecke (2013) exploits the introduction of a 'triple science' option in British high schools and show that those choosing this option were more likely to choose science courses in later grades. However, this effect was restricted to men. Justman and Mendez (2018) show that choosing STEM subjects in later high school years is not driven by prior differences in mathematical achievement, but that female students require stronger signals of mathematical ability to choose male-dominated subjects.

These results are very much in line with studies that focus on mathematical self-concept rather than on mathematical achievement. If mathematical self-concept differs between men and women, this will have implications for educational choices and later outcomes beyond differences in achievement. M. Jansen, Scherer and Schroeders (2015) find a relationship between the self-concept and the wish to study a STEM field. Perez, Cromley and Kaplan (2014) further show that graduating in STEM subjects with math-intensive courses can be explained by the self-concept. Watt and Eccles (2008) also link the self-concept to career choices.

A few studies have used quasi-experimental variation to study the effects of math curricula on later school outcomes and college major choices. Domina and Saldana (2012) examine the intensification of mathematics curricula in American high schools over the period from 1982 to 2004, suggesting that this intensification reduced social stratification in course credit completion but left inequality in more advanced subareas very pronounced. Cortes, Goodman and Nomi (2015) study an intensive math instruction policy that doubled instruction time for low-skilled 9th graders. They show that this policy had substantial positive effects on test scores, high school graduation, and college enrollment. Jia (2016) finds that stricter requirements increase STEM attainment to a certain extent, but only for white males. De Philippis (2017) also uses quasi-experimental variation in the form of a reform that allowed secondary schools in the U.K. to offer more science to high-ability 14-year-olds. Again, her results suggest that introducing this option increased men's willingness to enroll in STEM degrees but not women's.

Allensworth, Nomi, Montgomery and Lee (2009) study a reform that made a college preparatory curriculum in math and English mandatory. However, compared to this chapter, this reform referred to a lower school grade (9th grade). Further, the reform mainly aimed at an equalization between schools. Results are, therefore, not easily comparable to this chapter. The results reported in Allensworth et al. (2009) suggest almost no benefits for university freshmen or lower dropouts rates. Similarly, Jacob, Dynarski, Frank and Schneider (2017) use a statewide college-preparatory curriculum reform in Michigan, USA. They found that the increased emphasis on academic preparation in math and sciences increased the ACT science scores by 0.2 points. Although both studies looked at high school curriculum reforms, they did not study college subject choices or occupational decisions. Darolia, Koedel, Main, Ndashimye and Yan (2019) further show that for Missouri, USA, the differential access to high school courses in math or sciences does not affect higher education STEM enrollment or degree attainment. Their study, however, did not include a similarly extreme shock like the one exploited in this chapter.

A much smaller literature has focussed on the effects of math and science curricula on STEM choices and outcomes *outside the education system*. One strand of the literature started by Altonji (1995) considers the effects of math curricula on later wages (for example, Goodman, 2019; Joensen & Nielsen, 2009, 2016; Rose & Betts, 2004). However, these contributions typically do not address the question of STEM vs. non-STEM occupations. For example, Goodman (2019) shows that state changes in minimum high school math requirements substantially increased underprivileged students' completed math coursework and later earnings. Joensen and Nielsen (2016) and Joensen and Nielsen (2016) exploit a pilot scheme in Denmark that reduced the cost of choosing advanced math in ninth grade high school. Their results suggest that only females benefited from the pilot scheme in the form of a higher rate of completed STEM degrees and higher later earnings. Joensen and Nielsen (2009) and Joensen and Nielsen (2016) do not address whether more math in high school increases the inflow into STEM occupations as we do in this chapter. S. L. Morgan, Gelbgiser and Weeden (2013) examine college major selection and occupational *plans* but not actual outcomes. In fact, we are not aware of any other study that uses quasi-experimental variation in school curricula to study its effect on the actual participation in STEM occupations.

A more general literature has looked at possible sources of gender differences for different fields. Buser et al. (2017) and Gneezy, Niederle and Rustichini (2003) point out that an important factor behind the STEM gap may be gender differences in willingness to compete. Another possible source of female underrepresentation may be biased self-assessment (Correll, 2001), stereotypes and identity issues (Brenoe & Zölitz, 2020; Cech, Rubineau, Silbey & Seron, 2011; Cheryan, 2012; Del Carpio & Guadalupe, 2018; Franceschini et al., 2014). For example, Brenoe and Zölitz (2020) show that a higher share of females in a classroom in high school reduces the probability for women to study a STEM course. This is in line with Franceschini et al. (2014), who suggest that women are more easily intimidated by 'stereotype threat', i.e., by pieces of information that make STEM subjects appear inappropriate for them. Another possible source for gender differences in STEM participation are differences in mathematical self-concept (for example, Hübner et al., 2017).

Finally, we point out three studies that have examined the same reform as we study in this chapter but based on different data sets, different methods, and different research questions. Görlitz and Gravert (2016) and Görlitz and Gravert (2018) use aggregate time series provided by the statistical offices to estimate the effects of the reform on high school dropout and enrollment in higher education. Their results suggest increases in high school dropouts that dissipate in the medium run as well as higher enrollment into tertiary education, including STEM, but robustly so only for men. Görlitz and Gravert (2016) and Görlitz and Gravert (2018)'s papers do not use micro-data as we do in this chapter, consider only a few post-reform years and do not address long-term outcomes such as completed degrees or occupational choices. The study by Hübner et al. (2017) focus on the effects of the reform on high school achievement and university entry decision. Based on a before-after comparison using only data from the reform state Baden-Württemberg, they find that gender differences decreased for math achievement but increased for math self-efficacy (which is related to math self-concept). They do not find significant reform effects on initial study choices. A major difference between these studies and this chapter is that our data covers a much longer post-reform period allowing us to consider longer term outcomes such as the completion of STEM degrees (rather than their initial choice) and the entry into STEM occupations after successful graduation.

3.3 Institutional background

In the German education system, educational policies are largely determined at the federal state level, allowing states some degree of freedom to deviate from the general structure of the school system that is shared by all states. Using this freedom, the federal state of Baden-Württemberg (the third largest of all federal states) introduced a significant reform of the high school curriculum in 2002 which provides a natural experiment.[2]

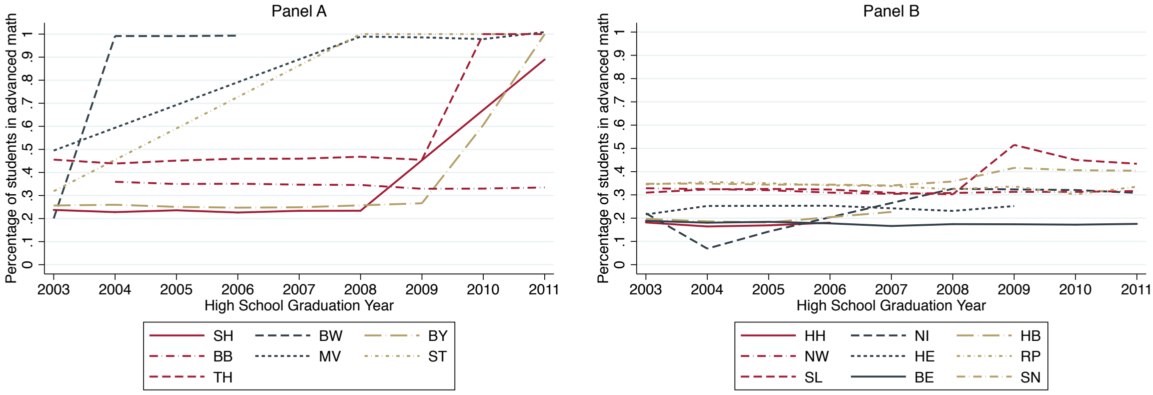
[2] Apart from Baden-Württemberg, only the federal states of Mecklenburg-Vorpommern and Sachsen-Anhalt carried out other school reforms at the high school level during the period of interest, which is why we drop these states from our analysis.

The German school system has a clear tracking structure in which students are allocated to one of three secondary school tracks after the fourth (or in some cases after the sixth) grade. Only the highest secondary track (*Gymnasium*, i.e., academic high school) grants an unconditional higher education entrance qualification (HEEQ). In addition to *Gymnasium*, there are more specialized/vocational high schools as well as a number of more indirect ways to obtain a HEEQ. However, these are chosen only by a minority of students.

The pre-reform high school curriculum was similar in all German states. In this curriculum, all students completed similar courses during the first ten or eleven school years. In grade ten or eleven, however, students were asked to choose a specific combination of subjects for the last two years of high school, with mild restrictions on which combinations and exam levels were possible. Out of the chosen classes, two had to be on an advanced level and several others on a basic level. In Baden-Württemberg, for example, at least one basic math and one basic German class had to be taken, in addition to at least one natural science class. If a math, German, or science class at an advanced level was chosen, students could fill their basic courses with other subjects. If they chose their math, German and science class as basic courses, however, they were free to choose other subjects, such as languages, arts or even sports as their advanced classes. Advanced classes were held five hours a week, basic classes only two hours per week. Given the nature of specialized/vocational high schools, the choices at these schools were less flexible. In both types of schools, three written high school exams and one oral exam in two advanced and two basic courses had to be passed to earn a HEEQ.

In 1999, Baden-Württemberg announced a reform affecting students starting their second to last school year in a *Gymnasium* from 2002 onwards. Specialized/vocational high schools introduced a modified version of these reforms one year later. As a consequence, academic high school graduates from 2004 onwards and specialized/vocational high school graduates were affected by the reform from 2005 onwards. The post-reform high school curriculum forced all students to attend a mandatory advanced class in mathematics, German, and one foreign language. In addition, two more advanced courses in one natural science and/or another foreign language had to be taken. This means that the total number of mandatory natural science courses increased from one to two (with both potentially at the basic level). Because of the larger number of required classes, advanced classes were reduced from 5 to 4 hours per week. The overall minimum instruction time per week increased from 26 hours per week to 30 hours. Apart from some additional aspects, it is fair to say that the essential content of the reform was a substantial shift towards more instruction time in math and the natural sciences for a large number of students who previously would not have chosen these subjects at all, or would have only chosen them at the basic level (and thus with only half the instruction time).

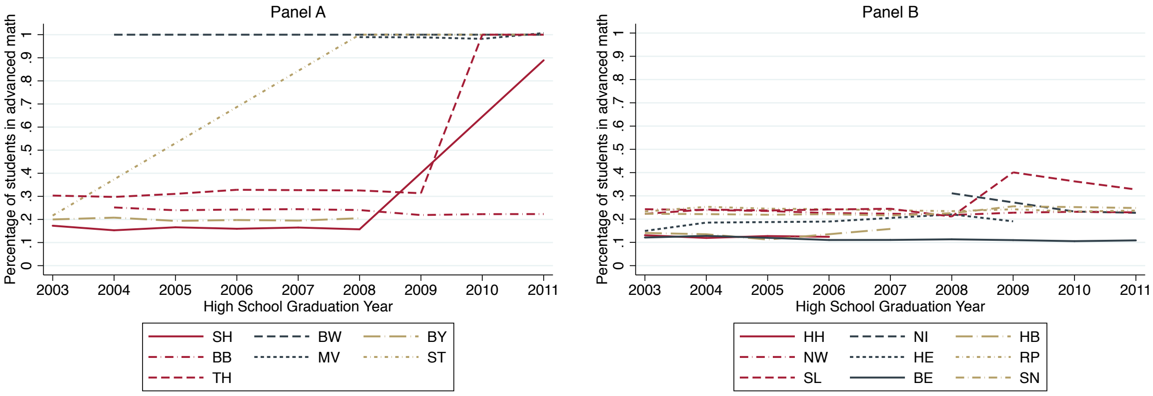
Figure 3.1 – Students taking advanced math per state



Note: Panel A includes the states Brandenburg (BB), Baden-Württemberg (BW), Bayern (BA), Mecklenburg-Vorpommern (MV), Schleswig-Holstein (SH), Sachsen-Anhalt (ST) and Thüringen (TH). Panel B shows the development for Saarland (SL) and Sachsen (SN), Berlin (BE), Bremen (HB), Hessen (HE), Hamburg (HH), Nordrhein-Westfalen (NW), Rheinland-Pfalz (RP) and Niedersachsen (NS). The data is provided by each state on a voluntary basis, leading unfortunately to some missing years. *Source:* Statistical Offices of the Federal States.

In order to illustrate the drastic nature and the comprehensive reach of the reform, Figures 3.1 and 3.2 present administrative data showing the impact of the reform compared to the situation in other federal states. The figures refer to the second qualification phase at the *Gymnasium*, i.e., grades eleven to thirteen. Administrative data is available from 2002 onwards, but a number of missing values and differences in coding make some values before 2003 unusable. As Figure 3.1 shows, advanced math participation in 2004

Figure 3.2 – Female students taking advanced math per state



Note: Same graph as above, only for females. Some states did not provide gender specific information for all years, which is why there are some missing years. *Source:* Statistical Offices of the Federal States.

varied from around .08 for Niedersachsen up to 1 for Baden-Württemberg. The graph demonstrates that the reform in Baden-Württemberg had a very substantial impact. Without mandatory advanced math classes, the highest share was around .5 in Saarland. All other states range between .1 and .35. Only Thüringen was constant above .4. Figure 3.2 shows that the proportion of females taking advanced math classes was generally lower than that of males. Again some values are missing, for example, because no gender-specific administrative data are available. Unfortunately, the value for Baden-Württemberg in 2003 is missing as well. We have no reason, however, to believe that the gender difference in Baden-Württemberg before the reform was much different from that in other states. The percentage of females taking non-mandatory advanced math classes ranged between .10 and .25 with a maximum of around .40. This difference shows that, in general, females were more affected by the reform than males.

Taken together, these numbers illustrate the dramatic impact the curriculum reform had on the level and instruction time for math during the last two years of high school. For over 80% of students, instruction time increased from three to four hours per week as students, who would have enrolled in a basic course before the reform (3 hours per week), were forced into an advanced course (4 hours a week). For women, the percentage of students receiving more instruction time was even higher as the share of female students taking advanced math courses before the reform was below that of men. For the natural sciences, the reform had a similarly strong impact on instruction time through the introduction of an additional advanced level course in one of the natural sciences (details not shown here). Another aspect of the reform was that it 'leveled the playing field' because it made participation in advanced math and natural sciences mandatory. In addition, it forced all students into a common classroom, exposing students to a wider spectrum of math abilities, as opposed to the situation before in which students were separated into basic and more advanced classes.

3.4 Econometric methods

3.4.1 Difference-in-differences estimation

We employ a difference-in-differences setup for our estimations with gender interactions in order to obtain gender-specific difference-in-differences effects. This setup compares the situation before and after the reform with the non-treated federal states serving as a counterfactual for the treated state. Our regression model is

$$\begin{aligned} y_{ist} = {} & \alpha + \gamma_1\, \mathit{After}_t + \gamma_2\, \mathit{BaWu}_s + \gamma_3\, \mathit{Female}_{ist} \\ & + \gamma_4\, (\mathit{After}_t \cdot \mathit{Female}_{ist}) + \gamma_5\, (\mathit{BaWu}_s \cdot \mathit{Female}_{ist}) \\ & + \rho\, \mathit{Treatment}_{ist} + \lambda\, (\mathit{Treatment}_{ist} \cdot \mathit{Female}_{ist}) \\ & + \boldsymbol{\beta}'\, \mathbf{X}_{ist} + \boldsymbol{\mu}'\, \mathbf{Z}_{st} + \eta_s + \nu_t + \epsilon_{ist} \end{aligned} \tag{3.1}$$

where the index s indicates in which federal state individual i obtained their higher education entrance qualification (HEEQ, high school degree, in Germany called *Abitur*) and t denotes the year the individual obtained their HEEQ. The dependent variable y_{ist} represents a binary outcome, for example, whether or not the individual i from state s

and HEEQ-year t later completed a STEM degree or worked in a STEM occupation. The dummies $After_t$ and $BaWu_s$ indicate whether the individual obtained her HEEQ after the reform year of 2004 (rather than before), and whether the individual received her HEEQ in the state of Baden-Württemberg (rather than elsewhere). The treatment variable $Treatment_{ist}$ is the product of $After_t$ and $BaWu_s$ and indicates whether the individual's high school curriculum was changed by the reform. The vector $\mathbf{X}_{ist}$ contains a number of individual covariates explained in more detail below, while $\mathbf{Z}_{st}$ represents time-varying state level characteristics controlling for time-varying differences between federal states. Finally, η_s and ν_t represent HEEQ-state and -year fixed effects.

In order to differentiate the difference-in-differences effects by gender, we include interactions of the difference-in-differences terms with a dummy $Female_{ist}$ indicating whether individual i is a woman.[3] As a consequence, ρ represents the treatment effect for men (i.e. for individuals with $Female_{ist} = 0$), while $\rho + \lambda$ represents the treatment effect for women (i.e. for individuals with $Female_{ist} = 1$). The parameter λ represents the gender difference in the reform effect. Overall, this setup identifies the reform effects ρ and $\rho + \lambda$ by comparing individuals before and after the reform in Baden Württemberg with the situation before and after the reform year in other federal states taken as a counterfactual scenario. There may be general time-constant differences between treatment and control states η_s as well as common time trends in STEM participation ν_t (common across all states). Moreover, we include into $\mathbf{X}_{ist}$ a large number of time-varying covariates at the state level (such as income per capita, unemployment rate, the density of tertiary institutions, see below) that aim to pick up potentially differential developments in STEM participation across states.

The reform effects ρ and $\rho + \lambda$ represent the total effects of the reform, i.e., those for the larger group of individuals who would have had much lower exposure to math and the natural sciences without the reform, and those for the smaller group of individuals who would have participated in advanced math and natural science courses anyway, but whose instruction time would have been slightly higher without the reform. Despite its

[3] We also carried out DiD regressions for both genders separately. These yielded very similar results but have the disadvantage of not directly showing the gender difference. Results are available on request.

mixed nature, the treatment effect estimated here represents an interesting and relevant policy parameter corresponding to a well-defined real-world intervention, which could, in principle, be implemented in other federal states as well.

In order to detect possible heterogeneity in the effects of the reform across ability levels, we augment Equation (3.1) by further interacting the treatment indicator and the gender difference with dummies for the high school GPA. For this purpose, we group the GPA into three categories and leave out the middle group as the reference group. In Germany, the High School GPA ranges from 0.7 (best) up to 4.0 (worst). Students with a GPA above 4.0 do not graduate. We grouped the best students with a High School GPA below two in $\textit{HS-GPA-1}_{ist}$ and above and equal to three in $\textit{HS-GPA-3}_{ist}$. This specification can be expressed as follows:

$$
\begin{aligned}
y_{ist} = {} & \alpha + \gamma_1 \; \textit{After}_t + \gamma_2 \; \textit{BaWu}_s + \gamma_3 \; \textit{Female}_{ist} \\
& + \gamma_4 \; (\textit{After}_t \cdot \textit{Female}_{ist}) + \gamma_5 \; (\textit{BaWu}_s \cdot \textit{Female}_{ist}) \\
& + \rho_1 \; \textit{Treatment}_{ist} \\
& + \rho_2 \; (Treatment_{ist} \cdot \textit{HS-GPA-1}_{ist}) \\
& + \rho_3 \; (Treatment_{ist} \cdot \textit{HS-GPA-3}_{ist}) \\
& + \lambda_1 \; (\textit{Treatment}_{ist} \cdot \textit{Female}_{ist}) \\
& + \lambda_2 \; (\textit{Treatment}_{ist} \cdot \textit{Female}_{ist} \cdot \textit{HS-GPA-1}_{ist}) \\
& + \lambda_3 \; (\textit{Treatment}_{ist} \cdot \textit{Female}_{ist} \cdot \textit{HS-GPA-3}_{ist}) \\
& + \boldsymbol{\beta}' \; \mathbf{X}_{ist} + \boldsymbol{\mu}' \; \mathbf{Z}_{st} + \eta_s + \nu_t + \epsilon_{ist} \qquad (3.2)
\end{aligned}
$$

3.4.2 Few treated clusters

It is well-known that in difference-in-differences setups, it is crucial to control for potential intra-cluster dependence. In our application, there is clustering in both the state and the time dimension. Ignoring intra-cluster dependence will bias standard errors downward and lead to over-rejection rates (Bertrand, Duflo & Mullainathan, 2004). As explained in the previous section, we include a large set of time-varying covariates at the state level as well as additional time and state effects in our difference-in-differences regressions to pick up potentially differential time trends across states. This will already take out a fair amount of intra-cluster correlations, mitigating potential problems of cluster inference.

Our application is characterized by a small number of clusters, of which only one is the treated cluster. For this and similar cases, Mackinnon and Webb (2017) compare the wild bootstrap, the wild cluster bootstrap, and an intermediate case called wild subcluster bootstrap. Their results suggest that for our scenario (one treated cluster, thirteen untreated clusters), the ordinary (= individual) wild bootstrap performs best.

Mackinnon and Webb (2017) also advocate comparing restricted and unrestricted bootstrapped p-values (i.e., with and without imposing the null hypothesis) as a diagnostic test for the validity of p-values. If the two coincide, this can be taken as an indication for their validity. Following this procedure, we found the ordinary (= individual) wild bootstrap to be the most adequate for our application. We, therefore, report p-values and confidence intervals based on the ordinary wild bootstrap (unrestricted version) throughout our results (given the chosen procedure, the results using the restricted version are similar and available on request).

3.5 Data

The data for this chapter were provided by the Centre for Higher Education Research and Science Studies (DZHW), see Baillet et al. (2017, 2019). The DZWH starts a new representative survey of German university graduates every four years. The survey includes rich information on parental background, the individual's higher education entrance qualification, choices during university study, and labor market entry. The main target population are all higher education graduates from institutions that are approved by the state. This includes universities as well as applied universities and similar institutions. The sample was drawn at the level of the institution using a stratified cluster sampling (Baillet et al., 2017). For our analysis, we use the cohorts 2005 and 2013. This provides a clear separation into four groups of university graduates who completed their final high school years in either the pre- or the post-reform period and in either the reform state Baden-Württemberg or other states. Table 3.1 summarizes the four groups.

Table 3.1 – Categorization of DiD groups for the analysis

Group	Before Treatment:	After Treatment:
Control:	HEEQ obtained before 2004 in control states	HEEQ obtained after 2004 in control states
Treatment:	HEEQ obtained before 2004 in Baden-Württemberg	HEEQ obtained in and after 2004 in Baden-Württemberg

Note: The table specifies the four categories needed for the empirical analysis. HEEQ: Higher education entrance qualification, i.e. high school graduation (*Abitur*). Note that the year of HEEQ is not necessarily the start enrollment at university. Cohort 2005 includes only individuals with HEEQ years between 1997 and 2001, cohort 2013 only from 2005 and 2009.

Note that each cohort includes students with different HEEQ years as study durations differ, and as students do not necessarily start their studies immediately after obtaining the higher education entrance qualification. The HEEQ year represents the year in which the higher education entrance qualification was obtained, not the year in which the person enrolled in tertiary education. In our analysis, we exclude individuals with a HEEQ obtained before 1997 and after 2001 for the 2005 cohort, and before 2005 and after 2009 for the 2013 cohort in order to drop unrepresentative long- and short-term students.

In this way, we also exclude high school graduates who might have experienced an announcement effect as well as the year 2004 in which only the *Gymnasium* implemented the reform but not certain other institutions that may also grant a higher education entrance qualification (*Fachschulen*).

Table 3.2 shows some basic sample information by gender. The two cohorts have approximately the same size. The individual-level covariates included in our difference-in-differences regressions are gender, age, parental education in four categories, parental occupation in two categories as well as state and year of the HEEQ. Table 3.2 further presents summary statistics for the degree and occupational outcome variables used in the regressions. The degree variables are dummies indicating whether or not a particular individual obtained a degree in a particular field. Labels such as 'at least one STEM degree' mean that we have a small number of individuals with more than one degree but count them as STEM if at least one of their degrees is in STEM. Following common practice, we include into STEM all fields in science, technology, engineering, and mathematics. More precisely, our STEM category includes the sciences (biology, chemistry, pharmacy, geosciences, physics), technology (computer science), engineering (all subfields of engineering), and mathematics. As indicated above, we also consider smaller subsets of STEM fields: mathematics and natural sciences (MatNat) and engineering and computer sciences (EngComp).

For occupational outcomes, our data include the KldB occupation code (German classification of occupations). For 2005, this is the KldB 1992, whereas for the other cohorts it is the KldB 2010. The German Federal Employment Agency provides a categorization into STEM and non-STEM occupations, but only for the KldB 2010 (Bundesagentur für Arbeit, 2019). For the KldB 1992 codes, we followed a translation from KldB 1992 to KldB 2010. This left us with a small number of cases for which it was not possible to assign a clear STEM or non-STEM status based on the 2010 STEM classification (because these occupations were more or less specific in the KldB 1992 classification than in the KldB 2010 classification). In order to resolve these cases, we employed a specific algorithm, the details of which are available on request.

Table 3.2 – Descriptive statistics

Variables	Males		Females	
	Mean	SD	Mean	SD
DiD				
Treated individuals	0.053	0.223	0.059	0.235
HEEQ after treatment	0.414	0.493	0.446	0.497
HEEQ in treatment state Baden-Württemberg	0.144	0.351	0.138	0.345
Age and Parents				
Age	26.545	1.796	25.919	1.701
Highest parental education: Other	0.014	0.116	0.011	0.104
Highest parental education: Vocational training	0.357	0.479	0.355	0.478
Highest parental education: HS Diploma	0.049	0.216	0.050	0.218
Highest parental education: PhD, Uni & AU	0.581	0.493	0.584	0.493
Highest parental occupation status: White collar	0.944	0.230	0.951	0.216
Highest parental occupation status: Blue collar and other	0.056	0.230	0.049	0.216
State variables, year of HEEQ				
Non working population per capita	0.047	0.023	0.047	0.024
Labor force participation per capita	0.502	0.025	0.503	0.025
Unemployment rate by gender	9.187	3.569	9.969	4.563
GDP per capita	27.485	6.633	27.847	6.948
Share of producing sector	0.087	0.029	0.086	0.030
Share of manufacturing sector	0.199	0.045	0.195	0.045
R&D per capita	0.092	0.045	0.096	0.048
Exports per capita	7.055	3.511	7.261	3.699
Imports per capita	6.619	4.321	6.912	4.589
Density of universities	2.067	0.454	2.071	0.473
Density of applied universities	3.501	1.092	3.578	1.159
Year	2002.094	4.374	2002.712	4.373
Mediators				
Finale grade of HEEQ	2.248	0.602	2.155	0.594
Other path than academic HS	0.140	0.347	0.080	0.271
Employment before university	0.249	0.432	0.246	0.430
Vocational training before university	0.145	0.352	0.114	0.318
Applied university	0.298	0.458	0.205	0.404
Degree type: teaching profession	0.047	0.212	0.146	0.353
Outcomes				
At least one degree in STEM	0.554	0.497	0.263	0.440
At least one degree in MatNat	0.130	0.336	0.145	0.352
At least one degree in EngComp	0.425	0.494	0.119	0.324
Current or last occupation in STEM	0.427	0.486	0.157	0.351
Current or last occupation in MatNat	0.025	0.151	0.028	0.159
Current or last occupation in EngComp	0.401	0.482	0.129	0.321

Note: HEEQ: Higher education entrance qualification. The two German states, Sachsen-Anhalt and Mecklenburg, are not included because they had a different reform during the period of interest. For all variables and the degree outcomes, we have 5199 male and 7652 female observations. For the occupation outcomes, we have 3664 male and 5470 female observations. For the regressions using the occupations as outcomes, we merge the state variables to the year of the degree. The German occupation classification *KldB* is used for classifying individuals into different fields of occupation. Information on the states of the respective HEEQs are included in the appendix.

In order to control for potential time-varying differences between federal states and in order to minimize remaining intra-cluster correlation, we also include a set of state- and time-specific variables, as shown in Table 3.2. All variables are measured at the state level. They are merged to the year of the HEEQ for the degree regressions and to the year of the degree for the occupation regressions.

In the last step, we include variables whose realization was after the reform and which may, therefore, have been mediators of reform effects. As these variables might have been affected by the reform, their inclusion should proceed with caution. However, we also ran our difference-in-differences regressions taking each of these variables as an outcome but did not find any significant reform effects on them. Note that by including these variables, and all other individual-level variables, we control for potential compositional differences in the population before and after the reform.

3.6 Empirical results

3.6.1 Main specifications

Tables 3.3 to 3.5 present the regression results for the gender-difference Equation (3.1). Specification (1) only includes the variables of the basic difference-in-differences equation (before/after treatment, treatment/control states, and their interaction). Specification (2) adds to this equation the personal characteristics age, age squared, parental education, and occupation (see Table 3.2) as well as HEEQ state and HEEQ year fixed effects. In specification (3), we include in addition the time-varying federal-state variables (also listed in Table 3.2). Finally, specification (4) adds personal characteristics that were formed after treatment and which may, therefore, be mediators of potential treatment effects on final degrees or occupations (for example, final grade of HEEQ, type of university, see second but last panel of Table 3.2).

For the outcome STEM degree, Table 3.3 panel A shows a negative male treatment effect of -5 to -10 percentage points, but all four coefficients are statistically insignificant. The gender difference, though negative, is insignificant as well. However, the resulting female treatment effect (the sum of the baseline male effect and the gender difference) is around

Table 3.3 – Gender difference regressions for degrees

	(1)	(2)	(3)	(4)
Panel A:	*Dependent variable*: Degree in STEM			
Treatment ($\hat{\rho}_{GD}$)	−0.052	−0.055	−0.099	−0.103
OWB p-values	{0.5137}	{0.4484}	{0.1698}	{0.1543}
Gender Difference ($\hat{\lambda}_{GD}$)	−0.096	−0.091	−0.083	−0.091
OWB p-values	{0.2389}	{0.2552}	{0.3147}	{0.2591}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.148***	−0.146***	−0.182***	−0.194**
OWB p-values	{0.0002}	{0.0003}	{0.0005}	{0.0012}
R^2	0.0890	0.1035	0.1052	0.1135
Panel B:	*Dependent variable*: Degree in MatNat			
Treatment ($\hat{\rho}_{GD}$)	−0.019	−0.027	−0.026	−0.022
OWB p-values	{0.6815}	{0.4859}	{0.6544}	{0.6982}
Gender Difference ($\hat{\lambda}_{GD}$)	−0.101*	−0.094*	−0.097*	−0.095^{+}
OWB p-values	{0.0435}	{0.0335}	{0.0439}	{0.0664}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.120***	−0.121***	−0.123*	−0.117*
OWB p-values	{0.0009}	{0.0005}	{0.0309}	{0.0237}
R^2	0.0069	0.0174	0.0192	0.0547
Panel C:	*Dependent variable*: Degree in EngComp			
Treatment ($\hat{\rho}_{GD}$)	−0.037	−0.032	−0.076	−0.084
OWB p-values	{0.6871}	{0.6783}	{0.3167}	{0.2591}
Gender Difference ($\hat{\lambda}_{GD}$)	0.008	0.006	0.017	0.008
OWB p-values	{0.9236}	{0.9405}	{0.8341}	{0.9267}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.029	−0.026	−0.059	−0.076^{+}
OWB p-values	{0.3291}	{0.3632}	{0.1356}	{0.0561}
R^2	0.1241	0.1388	0.1428	0.1994
Set of covariates				
After, *BaWu*, *female* and its interactions	Yes	Yes	Yes	Yes
Age and parents	No	Yes	Yes	Yes
State and year fixed effects	No	Yes	Yes	Yes
State variables	No	No	Yes	Yes
Mediators	No	No	No	Yes
Observations	12858	12858	12858	12858

Note: Ordinary wild bootstrap (OWB) p-values in curly parentheses, calculated using the Stata command *boottest*, see Roodman, MacKinnon, Nielsen and Webb (2019). The state variables are merged to the year of the HEEQ. The female treatment effect was computed as the sum of the male baseline treatment effect and the gender difference. The significance was tested with the help of the command *boottest*. The sets of control variables are the same for all three panels. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

-15 to -20 percentage points and statistically significant.

For the MatNat degree, there are small and statistically insignificant effects for men, whereas the gender difference and the resulting female treatment effect are negative and significant (around -11 percentage points). Finally, the estimates for the EngComp degree are negative but generally insignificant for men and women (panel C of Table 3.3). An exception is the full specification in column (4) for women which shows negative effects that are marginally significant. Altogether, the results for the degrees suggest that there are no significant reform effects for male STEM degrees, but significant negative effects for females that are driven by the math and natural sciences degrees but not by engineering or computer science.

In order to get an idea where the women who turned away from STEM degrees went to, Table 3.4 displays the corresponding results for other subject degrees. There are significant positive female reform effects on the completion of language and social science degrees but no such effects for men. This suggests that the reform deterred some women from completing STEM degrees, directing them to social sciences and languages.

The results for STEM occupations are shown in Table 3.5. For men, these effects are, in most cases, negative and small but statistically insignificant. The same is true for women, although there is a small and significant gender difference of -3 percentage points for MatNat occupations (column (4) of panel B). The general conclusion is that the reform did not change the fraction of men or women who work in STEM occupations.

How can these findings be rationalized? The first notable result is that there is no significant reform effect on men. Despite its substantial nature, the reform does not appear to have changed the share of male university graduates who complete a STEM degree (if anything, the effect of the reform on men was also negative). This is surprising given that the reform may, in principle, have changed both the preferences for STEM subjects and the preparedness for successfully completing a STEM degree. Note that, while a potential effect on STEM preparedness is unambiguously positive, an effect on preferences may be negative if the additional exposure to math and natural sciences deters individuals

Table 3.4 – Gender difference regressions for other subject degrees

	(1)	(2)	(3)	(4)
Panel A:	*Dependent variable*: Degree in Languages			
Treatment ($\hat{\rho}_{GD}$)	-0.012	-0.017	-0.018	-0.006
OWB p-values	{0.6350}	{0.5357}	{0.5404}	{0.7091}
Gender Difference ($\hat{\lambda}_{GD}$)	0.040	0.049	0.052	0.084^{*}
OWB p-values	{0.3421}	{0.2573}	{0.2913}	{0.0236}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	0.028	0.032	0.034	0.078^{**}
OWB p-values	{0.5949}	{0.4684}	{0.3195}	{0.0041}
R^2	0.0618	0.0731	0.0751	0.2081
Panel B:	*Dependent variable*: Degree in Social Sciences			
Treatment ($\hat{\rho}_{GD}$)	0.072	0.067	0.062	0.058
OWB p-values	{0.2423}	{0.2674}	{0.3101}	{0.3051}
Gender Difference ($\hat{\lambda}_{GD}$)	0.038	0.045	0.035	0.016
OWB p-values	{0.5198}	{0.4551}	{0.5707}	{0.8057}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	0.110^{*}	0.112^{*}	0.097^{*}	0.074^{+}
OWB p-values	{0.0145}	{0.0222}	{0.0496}	{0.0799}
R^2	0.0056	0.0112	0.0123	0.0624
Panel C:	*Dependent variable*: Degree in Medicine			
Treatment ($\hat{\rho}_{GD}$)	-0.005	0.001	0.042^{+}	0.042^{*}
OWB p-values	{0.7441}	{0.9292}	{0.0506}	{0.0142}
Gender Difference ($\hat{\lambda}_{GD}$)	0.011	-0.004	-0.006	-0.011
OWB p-values	{0.6421}	{0.8472}	{0.7512}	{0.6027}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	0.006	-0.003	0.036^{+}	0.031
OWB p-values	{0.8440}	{0.9075}	{0.0934}	{0.1089}
R^2	0.0094	0.0867	0.0916	0.1428
Set of covariates				
After, *BaWu*, *female* and its interactions	Yes	Yes	Yes	Yes
Age and parents	No	Yes	Yes	Yes
State and year fixed effects	No	Yes	Yes	Yes
State variables	No	No	Yes	Yes
Mediators	No	No	No	Yes
Observations	12858	12858	12858	12858

Note: Ordinary wild bootstrap (OWB) p-values in curly parentheses, calculated using the Stata command *boottest*, see Roodman, MacKinnon, Nielsen and Webb (2019). The state variables are merged to the year of the HEEQ. The female treatment effect was computed as the sum of the male baseline treatment effect and the gender difference. The significance was tested with the help of the command *boottest*. The sets of control variables are the same for all three panels. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.5 – Gender difference regressions for occupations

	(1)	(2)	(3)	(4)
Panel A:	*Dependent variable*: STEM occupation			
Treatment ($\hat{\rho}_{GD}$)	−0.011	−0.003	−0.048	−0.037
OWB p-values	{0.9105}	{0.9710}	{0.6817}	{0.6771}
Gender Difference ($\hat{\lambda}_{GD}$)	−0.016	−0.023	−0.033	−0.031
OWB p-values	{0.7741}	{0.7138}	{0.5580}	{0.6054}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.027	−0.026	−0.081	−0.068
OWB p-values	{0.6529}	{0.6368}	{0.3260}	{0.2584}
R^2	0.0974	0.1163	0.1211	0.2012
Panel B:	*Dependent variable*: MatNat occupation			
Treatment ($\hat{\rho}_{GD}$)	0.003	0.002	0.017	0.020
OWB p-values	{0.8882}	{0.9264}	{0.5262}	{0.4368}
Gender Difference ($\hat{\lambda}_{GD}$)	−0.024^{+}	−0.024	−0.024	−0.030*
OWB p-values	{0.0923}	{0.1178}	{0.1023}	{0.0416}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.021	−0.022	−0.007	−0.010
OWB p-values	{0.2519}	{0.1928}	{0.8120}	{0.7343}
R^2	0.0014	0.0069	0.0083	0.0148
Panel C:	*Dependent variable*: EngComp occupation			
Treatment ($\hat{\rho}_{GD}$)	−0.037	−0.032	−0.076	−0.084
OWB p-values	{0.8566}	{0.9380}	{0.6645}	{0.6228}
Gender Difference ($\hat{\lambda}_{GD}$)	0.008	0.001	−0.009	−0.001
OWB p-values	{0.8701}	{0.9811}	{0.8538}	{0.9790}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.029	−0.031	−0.085	−0.085
OWB p-values	{0.9101}	{0.9293}	{0.5475}	{0.4931}
R^2	0.1076	0.1237	0.1280	0.2116
Set of covariates				
After, *BaWu*, *female* and its interactions	Yes	Yes	Yes	Yes
Age and parents	No	Yes	Yes	Yes
State and year fixed effects	No	Yes	Yes	Yes
State variables	No	No	Yes	Yes
Mediators	No	No	No	Yes
Observations	9138	9138	9138	9138

Note: Ordinary wild bootstrap (OWB) p-values in curly parentheses, calculated using the Stata command *boottest*, see Roodman, MacKinnon, Nielsen and Webb (2019). The state variables are merged to the year of the HEEQ. The female treatment effect was computed as the sum of the male baseline treatment effect and the gender difference. The significance was tested with the help of the command *boottest*. The sets of control variables are the same for all three panels. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

from studying these subjects in an even more intensive way at university.

This is also a possible explanation for the negative reform effect found for females. The additional math content may have deterred some females from choosing STEM majors at university either by changing their preferences or by changing other factors that determine the choice of STEM subjects. Plausible mechanisms include an effect on math self-concept and peer effects resulting from the fact that the reform stopped the separation into groups following basic vs. advanced math courses. This explanation is in line with Hübner et al. (2017), who found that female students experienced a lowered mathematical self-concept as a consequence of the reform. It is also known that females are more vulnerable to stereotype-threat, which is more likely to arise in an advanced math environment (for example, Franceschini et al., 2014). A more advanced math environment may also signal a higher level of competition in STEM subjects compared to other subjects, potentially putting off females (for example, Buser et al., 2017; Gneezy et al., 2003). If one assumes that having attended an advanced math or natural science class is a prerequisite for studying a STEM subject at university, then our results are also in line with Brenoe and Zölitz (2020) who found that a higher share of females in a classroom increases gender-stereotypical behaviors. Given that *all* women had to be present in advanced math classes, this might have therefore reduced the STEM orientation even of those women who would have been interested in studying a STEM subject before the reform.

Another possible channel of the reform may have been a higher dropout rate in high school, reducing the number of female STEM students at university. Görlitz and Gravert (2016) point to this effect of the reform based on administrative data time series. However, it seems unlikely to us that especially students interested in STEM subjects did not complete high school due to the reform as these students are typically of high ability and would probably not have failed the high school degree as a consequence of the higher math and science standards.

Our results indicate a negative effect of the reform on the share of female graduates completing a STEM degree but no change for the eventual share of women working in

STEM occupations after graduation. This suggests that those women who did not pursue STEM degrees because of the reform would not have worked in STEM occupations anyway despite having a STEM degree. This is a plausible scenario as the women whose behavior was changed by the reform are likely to have a more marginal interest in STEM subjects compared to those women whose behavior was not changed by the reform.

3.6.2 Heterogeneous reform effects

In this subsection, we consider potential heterogeneity in reform effects. Note that, even if the total effect of the reform is zero, there might be effects on population subgroups that compensate each other. Our results for heterogenous reform effects are shown in Tables B2 and B3, following Equation (3.2).

Table B2 for the degrees suggests that, for male STEM and EngComp degrees, the negative baseline coefficient for the large group of average ability students (GPA-2) is arithmetically counteracted by a positive coefficient for low ability students (GPA-3). This would suggest that the reform only affected average and high ability students. However, none of these effects is statistically significant. For women, we observe a similar pattern, i.e., the (significant) negative effects are partly counteracted by positive coefficients for low ability students. This also suggests that the reform mainly affected average and high but not low ability students. However, again, the differential effects are generally not statistically significant.

The corresponding results for occupations are shown in Table B3. Here, we observe similar patterns for men, i.e., negative baseline effects that are counteracted by positive coefficients for the low ability group. For men, these differential effects for the low ability group are even statistically significant, suggesting that low ability students were either not or even positively affected by the reform. For women, we observe no such patterns for occupations.

Overall, this presents weak evidence that, if there were negative reform effects, these mostly affected average and high ability students. However, given the low precision of these estimates, we view this conclusion as tentative.

3.7 Conclusion

This chapter analyzed the consequences of a substantial curriculum reform of the last two years of high school in one of the German federal states (Baden-Württemberg) on the share of male and female students who complete university degrees in STEM subjects and who work in STEM occupations after graduation. Our results suggest that, despite its drastic nature, the reform did not change the share of male graduates completing STEM degrees. Our analysis of effect heterogeneity indicates that there may have been opposing reform effects on low and high ability students. For women, we do find *negative* reform effects on the share of female STEM graduates, which are mainly driven by the subjects of math and natural sciences. It appears that the reform redirected some women from STEM subjects to languages and social sciences. Our interpretation of these results is that the fact that the reform forced women into advanced math and natural science courses that were compulsory for everybody may have had negative impacts on female math self-concept (Hübner et al., 2017) or implied peer effects that led to more gender-stereotypical behavior (Brenoe & Zölitz, 2020).

Our results further suggest that, although we observe significant negative effects on the completion of STEM degrees, the reform did not change the share of male or female individuals who later work in STEM occupations. This indicates that those who were deterred from pursuing STEM studies by the reform would not have worked in STEM occupations anyway. Overall, the results from the natural experiment considered by us suggest that it will be hard to increase STEM participation in the labor market, even if drastic changes in high school curricula are implemented. Future research should address in more detail potential mechanisms at play and examine whether earlier curricula interventions may be more influential.

Appendix

B.1 Tables

Table B1 – Addition to Table 3.2: Descriptive statistics on the states of HEEQ

	Males		Females	
Variables	Mean	SD	Mean	SD
States of HEEQ				
Schleswig-Holstein	0.022	0.146	0.025	0.156
Hamburg	0.020	0.140	0.021	0.145
Niedersachsen	0.116	0.320	0.113	0.316
Bremen	0.012	0.108	0.018	0.133
Nordrhein-Westfalen	0.139	0.346	0.142	0.349
Hessen	0.061	0.240	0.061	0.238
Rheinland-Pfalz	0.045	0.207	0.040	0.196
Bayern	0.229	0.420	0.220	0.414
Saarland	0.005	0.071	0.006	0.076
Berlin	0.025	0.157	0.027	0.162
Brandenburg	0.029	0.167	0.033	0.179
Sachen	0.093	0.291	0.102	0.302
Thüringen	0.060	0.237	0.055	0.229

Note: HEEQ: Higher education entrance qualification. The two German states, Sachsen-Anhalt and Mecklenburg, are not included because they had a different reform during the period of interest. The table is supplementary to Table 3.2.

Table B2 – Treatment effect heterogeneity along high school GPA for STEM degrees

Degree in	(1) STEM	(2) MatNat	(3) EngComp
Treatment ($\hat{\rho}_{GD}$)	−0.127	−0.012	−0.117
OWB p-values	{0.1613}	{0.8520}	{0.1987}
Treatment×GPA$_1$	−0.130	−0.039	−0.091
OWB p-values	{0.3871}	{0.5863}	{0.5145}
Treatment×GPA$_3$	0.143	−0.014	0.157
OWB p-values	{0.1462}	{0.7492}	{0.1217}
Gender Difference ($\hat{\lambda}_{GD}$)	−0.090	−0.098*	0.008
OWB p-values	{0.3383}	{0.0975}	{0.9287}
Gender Difference×GPA$_1$	0.124	−0.017	0.141
OWB p-values	{0.4148}	{0.8587}	{0.3226}
Gender Difference×GPA$_3$	−0.043	0.028	−0.061
OWB p-values	{0.3843}	{0.6260}	{0.2142}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.217**	−0.110^{+}	−0.109*
OWB p-values	{0.0010}	{0.0529}	{0.0135}
Female Treatment ×GPA$_1$($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.005	−0.056	0.050
OWB p-values	{0.9477}	{0.1224}	{0.3214}
Female Treatment ×GPA$_3$($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	0.100	0.014	0.096
OWB p-values	{0.2273}	{0.6316}	{0.1643}
Observations	12858	12858	12858
R^2	0.1144	0.0549	0.2005

Note: Ordinary wild bootstrap (OWB) p-values in curly parentheses, calculated using the Stata command *boottest*, see Roodman, MacKinnon, Nielsen and Webb (2019). The state variables are merged to the year of the HEEQ. The female treatment effect was computed as the sum of the male baseline treatment effect and the gender difference. The significance was tested with the help of the command *boottest.* $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B3 – Treatment effect heterogeneity along high school GPA for STEM occupations

	(1)	(2)	(3)
Occupation in	STEM	MatNat	EngComp
Treatment	−0.113	0.027	−0.140
OWB p-values	{0.2853}	{0.3558}	{0.3026}
Treatment×GPA$_1$	−0.055	−0.036	−0.018
OWB p-values	{0.6252}	{0.2022}	{0.8752}
Treatment×GPA$_3$	0.264*	−0.012	0.276^{+}
OWB p-values	{0.0348}	{0.6253}	{0.0592}
Gender Difference	0.045	−0.032^{+}	0.078
OWB p-values	{0.5413}	{0.0617}	{0.2512}
Gender Difference×GPA$_1$	0.084	0.040	0.044
OWB p-values	{0.3921}	{0.4009}	{0.6625}
Gender Difference×GPA$_3$	−0.290^{+}	−0.011	−0.279
OWB p-values	{0.0578}	{0.7453}	{0.1030}
Female Treatment ($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.0673	−0.0050	−0.0623
OWB p-values	{0.2567}	{0.8729}	{0.4736}
Female Treatment ×GPA$_1$($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	0.0295	0.0036	0.0258
OWB p-values	{0.6095}	{0.8882}	{0.5606}
Female Treatment ×GPA$_3$($\hat{\rho}_{GD} + \hat{\lambda}_{GD}$)	−0.0255	−0.0228^{+}	−0.0027
OWB p-values	{0.6398}	{0.0662}	{0.9604}
Observations	9138	9138	9138
R^2	0.2027	0.0150	0.2132

Note: Ordinary wild bootstrap (OWB) p-values in curly parentheses, calculated using the Stata command *boottest*, see Roodman, MacKinnon, Nielsen and Webb (2019). The state variables are merged to the year of the HEEQ. The female treatment effect was computed as the sum of the male baseline treatment effect and the gender difference. The significance was tested with the help of the command *boottest*. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 4

Impact of universities in a flat hierarchy: Do degrees from top universities lead to a higher wage?*

*This chapter is based on: Schwerter, J. (2020): Impact of universities in a flat hierarchy: Do degrees from top universities lead to a higher wage?, unpublished manuscript, University of Tübingen.

4.1 Introduction

For several countries, there is a rich set of empirical results showing that the decision to enroll at a given university is important for the wage after graduation. The literature almost unambiguously finds a positive wage premium for (subgroups of) graduates from universities with an elite status (for example, Andrews, Li & Lovenheim, 2016; Anelli, 2016; Birch, Li & Miller, 2009; Brand & Halaby, 2006; Brewer, Eide & Ehrenberg, 1999; Carroll, 2014; Carroll, Heaton & Tani, 2018; Hoekstra, 2009), a higher quality (for example, Black & Smith, 2004, 2006; Hussain, McNally & Telhaj, 2009; Jung & Lee, 2016; Long, 2008; Thomas & Zhang, 2005; Weinstein, 2017) or a high student selectivity (for example, W. Chen, Grove & Hussey, 2012; Dale & Krueger, 2002, 2014; Lindahl & Regnér, 2005; Milla, 2017; Monks, 2000; Thomas, 2003; Walker & Zhu, 2017). Most of the literature relies on findings from the United States, England, or Australia, which have a publicly known hierarchy of universities. It is unclear if this wage premium is also present in countries with a rather flat university hierarchy in which top universities are field-specific. To answer this question, I exploit graduate survey data from Germany. To the best of my knowledge, no other paper analyzes this wage premium for Germany so far.

To measure the quality of a university, I rely on two different rankings. The first one is the Quacquarelli Symonds World University (QS) Ranking. The QS is an international top university ranking, available since 2014. It is subject-specific and includes 50 to 500 universities, depending on the year and subject. Similar to other university rankings, they rank several standard university quality measures used in the literature. I extracted all of the German universities which were listed by these rankings to get a measure for top (or at least high-ranked) universities in Germany.

Another measure for "better" universities follows Avery, Glickman, Hoxby and Metrick (2013). They calculate a revealed preference ranking based on top students' university decisions. Top students are assumed to be free in their enrollment decision and thus collectively decide which universities are the best. The revealed preference of these students and the university's acceptance yield another ranking of universities through the joint decision of students and universities. Thus, I construct one ranking based on the mean high school GPA of students per cohort, university, and area of study to get a second subject-specific university ranking for Germany.

To overcome the possible selection problem due to simple OLS regression, I rely on an IV approach. The instrument in use is the number of top universities per state, cohort, and area of study. IV regression results indicate a wage premium of 11 to 13% for the QS ranking and 5 to 8% for the revealed preferences and acceptance (RPA) ranking. The QS ranking effect is mainly prevalent one year after graduation, while the RPA ranking effect remains stable even five years after graduation. Gender-specific regressions reveal that women are the main beneficiaries of a degree from a top university for both rankings. Against the background of a non-negligible gender wage gap females face in Germany, this is highly interesting.

The chapter continues as follows. Section 4.2 reviews the literature and Section 4.3 gives an insight into the data. Subsequently, the econometric model is presented in Section 4.4. After that, I present the estimation results in Section 4.5 and conclude in Section 4.6.

4.2 Literature

There are only a few studies that used university rankings to identify top universities' wage premia. Hartog, Sun and Ding (2010) are one of the few ever using a ranking to find a university wage premium. They use data on Chinese graduates and find a premium for graduates of the top 100 universities of 28% compared to those of the ranks 401-500. They use the ranking from the China University Alumni Association, which includes measures for research quality, quality of education, and reputation. Birch et al. (2009), Carroll (2014), and Carroll et al. (2018) use the ShanghaiRanking to identify the so-called *Group of Eight* universities in Australia.[1] While Birch et al. (2009) could not find any wage premium, Carroll (2014) and Carroll et al. (2018) find a small but statistically significant hourly wage premium of around 3 to 5%. For Italy, using a ranking based on admission information, Anelli (2016) finds a yearly income premium of 52%.

Apart from these studies, the literature relies on elite status, the selectivity of universities, and single quality measures. Brewer et al. (1999) report a substantial hourly wage premium of 14 to 30% for graduates of elite private colleges in the US relative to public colleges. Brand and Halaby (2006) find that graduates from an elite college have an advantage in educational achievement and occupational status. Results for wages are, however, mixed. They do find a long-run hourly wage premium of around 18.5% for graduates from elite universities, but no immediate effect. Similar to the idea of elite universities, Hoekstra (2009) analyzes the effect of so-called "flagship universities" of the US. Using regression discontinuity, he finds an earnings premium of around 24%, but only for white men. Jung and Lee (2016) rely on an official hierarchical classification of Korean universities and find that university prestige is vital for the wages of graduates, with an estimated wage premium of around 20%. They further show that results are more pronounced for males than for females.

Another strand of the literature relies on quality measures of a university rather than the status. Typical measures of quality are, for example, the mean test score, faculty-student ratio, retention rate, total tariff score, mean faculty salary, or expenditure per pupil. Long (2008) uses the average quality of universities within a certain radius of the students' location during high school. Across all methods of estimation, Long (2008) finds robust evidence of the positive effects of college quality on college graduation and household income and weaker evidence regarding hourly wages. The hourly earnings, for instance, increase for men by between 13 to 22%, while there is no statistically significant premium for women. Hussain et al. (2009) emphasize the positive impact of university quality on earnings of around 6% for a one standard deviation increase in university quality. They further conclude that the relationship is highly non-linear and that the top students

[1] Australian universities ranked in the top 100 in the world in 2012 according to the Academic Ranking of World Universities (ARWU) (later called ShanghaiRanking) are coded as 1 and 0 otherwise.

benefit the most, which is also a result in Thomas and Zhang (2005). Weinstein (2017) finds a positive effect of relative and absolute university quality on earnings one year after graduation but no long-run effects (i.e., ten years after graduation).

Another branch of the literature focuses on the selectivity of universities only, which means comparing more selective universities to less selective ones. Monks (2000) shows that graduates from highly selective universities earn about 8 to 13% more in hourly wages than those from less-selective institutes. They further show that the results are stable for men and women but are only statistically significant for white males and females. Dale and Krueger (2002) estimate the payoff of attending more selective universities and, therefore, match students who applied to the same colleges but got accepted differently to reduce the selection bias. Using this matching method, they find a wage premium only for the more selective universities for children from low-income families, but no general effect for everyone. Those results are, however, restricted to top tier elite schools. Dale and Krueger (2014) expanded the earlier study using administrative data, increasing the number of universities and still find only subgroup effects of university characteristics for blacks, Hispanics, and graduates from households with low educational background. They use average SAT scores, Basson's index of college selectivity, and net tuition as quality measures (instead of selectivity).

W. Chen et al. (2012) follow the matching method used by Dale and Krueger (2002) and find substantial results for more selective MBA programs of about 16% higher hourly wages, not just for specific subgroups. Ge, Isaac and Miller (2018) follow Dale and Krueger (2002) as well and confirm the missing significance for annual earnings only for males after controlling for the selection into highly selective universities. They expand the literature by showing that for women, there are indeed significant results. On the one hand, a highly selective university increases women's earnings as well as the probability of getting an advanced degree. On the other hand, it reduces women's likelihood of marriage, and the increase in earnings is higher for married women than for singles. Walker and Zhu (2017) match mean standardized admission scores for each field of study of an institution per cohort and further include the selectivity of each subject of the institute. They find a real gross hourly wage premium of 10% for males and 11% for females.

4.3 Data

This chapter's main interest is to analyze the impact of top universities on wages in a flat university hierarchy. For this purpose, I measure the university quality of the area of study. More specifically, I examine the effect of top programs within universities compared to the rest of the area of study. I use two different rankings to identify a university as being "better" than others. The first relies on the international, subject-specific Quacquarelli Symonds World University Rankings (QS). For the second ranking, I calculate a revealed preferences and acceptance (RPA) ranking based on the mean high school GPA of university graduates.

The QS ranks universities since 2014 for several areas of study. The number of universities in the listed sub-rankings differs from area to area and vary by year. Table C2 summarizes the areas of study and the number of universities ranked in a given year. The ranking includes information on (i) academic reputation, (ii) employer reputation, (iii) faculty-student ratio, (iv) citations per faculty, (v) the international faculty ratio and (vi) the international student ratio. These factors are weighted differently depending on the specific areas to end up with one ranking. The QS, however, does not report the specific weighting formulas.

The customized German QS ranking for the analysis is calculated as follows: first, I only keep all German universities from the rankings and then calculate the mean from 2014 to 2017 to get a more robust indicator for each area of study. Then, I convert the means into percentiles per area of study because the number of universities offering an area of study differs depending on the latter. Data confidentiality forbids the identification of universities of each individual. To overcome this problem, I take means of the percentiles of the two universities next to each other within the ranking, beginning from the top.[2] The identification of one particular university is thereby not possible anymore, but because the groups are based on the ranking, the signal should not change significantly in the regression. The better university of the remaining two is downgraded a little bit, and the worse one is upgraded. Assuming that the signal coming from a high-quality university decreases with the percentile ranking in the first place, estimates should be downward biased. If universities are not listed by the QS, they are not good enough to rank. One alternative specification of the ranking variable is a dummy variable indicating if a university is ever ranked for any study by the QS or not. The binary coded rank variables is a condensed version of the rank information, which may reflect the effects of the rank more directly. The other specification indicates an upper quantile dummy of the ranking.

For the revealed preferences and acceptance (RPA) ranking, I calculate the mean of students' high school GPA per university, area of study, and cohort. Then, I rank the universities based on the mean and calculate a ranking in percentiles based on the number of universities providing the area of study. The idea behind the RPA ranking is the following: students can decide in which university they want to enroll and thereby reveal what the best (possible) university is for them. If students do not have the best grades, they are bounded by the acceptance of universities that can select students. Thus, the ranking shows not only the revealed preferences of students but the best possible or accepted preference. The combination of all students' collective decision and the acceptance of the universities then results in a ranking of universities. Again, I convert the ranking into percentiles for a better comparison between the different areas of studies.

I am not using the ShanghaiRanking like Birch et al. (2009), Carroll (2014), and Carroll et al. (2018) because this ranking includes fewer areas of study. Further, the differences

[2] In case two universities have the same mean value they also have the same ranking. The next best university gets the very next number, not omitting any integers. Thereby, the ranking is condensed. Due to universities' grouping in groups of two, the resulting ranking percentiles are more coarse than the original ranking.

between the QS and ShanghaiRanking for German universities are very small. The German *CHE* is another ranking in which universities and areas of study can decide to participate or not. This ranking is less suitable because, in each area, numerous universities choose not to participate.

4.3.1 Sample

The analysis exploits the Graduate Panel of the German Centre for Higher Education Research and Science Studies (DZHW). Due to the sample and survey design, the DZHW Graduate Panel offers the best opportunities to comprehensively examine research questions about German university graduates (Baillet et al., 2017, 2019). The data is based on individuals graduating in 2005 and 2009, observed in two waves each. The first survey is conducted up to a year after graduation, the second after four to five years. Thus, I observe individuals in 2005/2006 and 2009/2010 in their first wave and 2010 and 2014 in their respective second wave. Moreover, I include only full-time employed individuals with just one degree. This latter restriction is important because in case students have two or more diplomas, it is unclear which is more important for hiring.

Note that since the Bologna reform (which changed the general degree-system from the German “Diplom” to the bachelor’s and master’s system) took place in-between 2006 and 2010, there are almost no graduates with a bachelor’s and a master’s degree in the sample, i.e. graduates with curricula based on two degrees. For the observed cohorts, the most typical degree (60% of the cases) is the German “Diplom” (Table 4.1). The “Diplom” comprises the bachelor’s and master’s degree. Thus, this group is unaffected by the restriction of just on degree mentioned earlier. Other “single” degrees are the state exam (11%) and the teaching degree (12%), which used to be common for these cohorts.

The complete set of variables used in the analysis is shown in Table 4.1 for three different (sub-)samples. The first three columns show general sample information, containing up to 16,453 full-time employed individuals with just one degree. Variables with less observations in column (1) are due to missing information. When keeping only individuals without any missing values, there are 10,218 observations left. Column (4) and (5) show the mean and standard deviation for this subsample of complete cases, here called RPA subsample. It is named RPA subsample because relying just on the RPA ranking, I can use all these observations in a regression. For the QS ranking, the number of observations drops to 6,573 because the QS ranks only a subset of areas. The mean and standard deviation are reported in columns (6) and (7).

Comparing the three subgroups mentioned above, the differences of the subsample means are low. The share of females is slightly affected by the QS ranking. The main reason for this is that some popular areas among women, such as German language studies, are not within the QS ranking. The complete sample has a female share of 0.59, the RPA subsample of 0.58, and the QS subsample of 0.52. This difference driven by the area of study decisions should be kept in mind when interpreting the results. Apart from this, for example, the age, the wave dummy, the cohort dummy, the high school GPA, high school graduation from a vocational school, attending an applied university (*Fachhochschule*), having been employed before university, and the dummy indicating children have very similar means. Thus, apart from the small gender difference, there does not seem to be a high selection due to the dropped observations.

Table 4.1 – Summary statistics

	Obs	Mean	SD	Mean	SD	Mean	SD
	General sample			RPA subsample		QS subsample	
Outcome							
Log monthly gross wage	15076	7.86	0.52	7.87	0.50	7.95	0.46
QS							
Customized QS ranking	10745	24.46	37.98	-	-	24.41	37.94
Indicator for QS ranked university	10745	0.30	0.46	-	-	0.30	0.46
Indicator for QS ranked and above 75th PCTL	10709	0.23	0.42	-	-	0.23	0.42
Indicator for QS ranked and above 90th PCTL	10709	0.15	0.36	-	-	0.15	0.36
Number of QS ranked universities per state, cohort and area	16412	0.62	0.94	-	-	0.95	1.01
Number of QS ranked universities above 75 PCTL per state, cohort and area	16412	0.52	0.82	-	-	0.81	0.90

Number of QS ranked universities above 90 PCTL per state, cohort and area	16412	0.40	0.72	-	-	0.60	0.83
RPA							
Mean HS GPA per uni, study and cohort	16448	2.28	0.33	2.28	0.33	2.26	0.33
Revealed preferences and acceptance ranking	16448	56.74	29.08	56.57	29.15	55.81	28.85
Indicator for RPA ranking above 75th PCTL	16405	0.33	0.47	0.32	0.47	0.30	0.46
Indicator for RPA ranking above 90th PCTL	16405	0.15	0.35	0.14	0.35	0.12	0.33
Number of RPA universities per state, area and cohort above 75th PCTL	16405	1.24	1.23	1.24	1.24	1.37	1.33
Number of RPA universities per state, area and cohort above 90th PCTL	16405	0.63	0.87	0.63	0.87	0.70	0.92
Basic							
Female dummy	16451	0.55	0.50	0.54	0.50	0.49	0.50
Age	16453	29.55	3.31	29.54	3.33	29.66	3.26
Wave	16453	0.39	0.49	0.37	0.48	0.37	0.48
Cohort	16453	0.40	0.49	0.38	0.48	0.38	0.48
Educational Background							
High school GPA	16261	2.27	0.61	2.27	0.61	2.25	0.62
Year of HEEQ	16346	1999.80	3.08	1999.64	3.06	1999.55	3.00
Field-specific HEEQ	16363	0.02	0.16	0.02	0.14	0.02	0.14
HEEQ from vocational school	16363	0.13	0.33	0.13	0.33	0.14	0.34
Foreign HEEQ	16363	0.01	0.09	0.00	0.02	0.00	0.02
High school at vocational school	16390	0.06	0.23	0.05	0.22	0.06	0.23
University							
Grade of University degree	15441	1.87	0.54	1.86	0.54	1.88	0.55
Type of degree: Magister	16453	0.05	0.22	0.06	0.23	0.04	0.20
Type of degree: Bachelor	16453	0.12	0.32	0.10	0.30	0.10	0.30
Type of degree: State Examination	16453	0.11	0.31	0.08	0.27	0.09	0.28
Type of degree: Teaching degree	16453	0.12	0.32	0.11	0.31	0.08	0.27
Type of degree: Other	16453	0.00	0.05	0.00	0.05	0.00	0.01
General University	16453	0.63	0.48	0.62	0.48	0.64	0.48
Experience before graduation							
Vocational training before university	16408	0.28	0.45	0.29	0.45	0.31	0.46
Employment before university	16391	0.31	0.46	0.32	0.47	0.32	0.47
Voluntary internship	16232	0.38	0.48	0.39	0.49	0.38	0.48
Mandatory internship	16420	0.55	0.50	0.53	0.50	0.51	0.50
Student assistant	16404	0.33	0.47	0.34	0.47	0.36	0.48
Working student	16404	0.34	0.47	0.35	0.48	0.37	0.48
Family Information							
Married	16328	0.21	0.41	0.20	0.40	0.20	0.40
Married and female	16326	0.11	0.31	0.10	0.30	0.09	0.29

Child-dummy	16315	0.15	0.35	0.14	0.35	0.14	0.35
Children and female	16313	0.07	0.25	0.06	0.24	0.06	0.23
State variables							
GDP per capita	15246	31.64	7.51	31.37	7.54	31.65	7.54
Imports per capita	15246	9.02	6.04	8.89	6.09	8.99	6.11
Exports per capita	15246	9.09	4.05	8.87	4.02	8.96	4.04
R&D expenses from businesses per capita	15246	0.59	0.36	0.57	0.35	0.58	0.35
R&D expenses from the states per capita	15246	0.13	0.07	0.13	0.07	0.13	0.07
Patents per capita	15246	0.63	0.42	0.60	0.41	0.61	0.41
Producing sector in percentage	15246	0.07	0.01	0.07	0.02	0.07	0.02
Manufacturing sector in percentage	15246	0.18	0.05	0.18	0.05	0.18	0.05
Density of universities	15246	2.26	0.57	2.29	0.57	2.30	0.57
NUTS-2 Variables							
Unemployment rate	12420	7.18	3.16	7.18	3.17	7.09	3.10
Income	12420	187.47	21.65	187.67	21.65	188.50	21.64

Note: The table shows summary statistics for all covariates used in the empirical analysis, pooled over both survey waves. The first column of observations only excludes individuals without a full-time job. The second column of observations contains only individuals for which all information in the RPA regression are available. The third further drops individuals with areas not ranked by the QS. The RPA-subsample includes 10218 individuals whereas the QS-subsample includes only 6573 individuals. PCTL: percentile. Summary statistics for the different areas of study and the federal states are included in Table C1 in the appendix. There are no entries for the QS ranking for the RPA subsample because the QS does not rank some subjects which are included in the RPA sample. *Source*: DZHW Graduate Panel 2005 and 2009, own calculations.

4.4 Econometric model

I use a general pooled OLS for the estimations to obtain baseline results. My regression model is

$$y_{it} = \alpha_{OLS} + \rho_{OLS}\ ranking_{it} + \boldsymbol{\beta}'_{OLS}\ \mathbf{X}_{it} + \epsilon_{it}$$

where i stands for the individual and t for the specific wave. The outcome variable y is the log of the current monthly gross wage in the first and second wave. The primary variable of interest is *ranking*, which will be either the QS ranking or the RPA ranking or a dummy for the top universities based on these rankings. The coefficient ρ_{OLS} measures the effect of a graduate from a top university. In $\mathbf{X}_{it}$, I include the variables listed in Tables 4.1 and C1. I control for basic wage regression variables, such as gender, age and age squared, the cohort and the wave, area of study, educational background, university decision information, and the final GPA, (work) experience before and during the studies, and family information. The vector of coefficients $\boldsymbol{\beta}'_{OLS}$ captures the effects of the respective variables. The general intercept is included with α_{OLS}, and ϵ_{it} is the idiosyncratic error term.

Although I employ a rich set of control variables, there might still be unobserved variables leading to biased estimates. The choice of a highly ranked university may be correlated with other wage relevant characteristics. To solve the endogeneity problem, I run instrumental variable regressions. The instrument for both rankings is the number of top universities in the state of the individuals' higher education entrance qualification. Thereby, I assume that students first choose an area of study and then the location of the university.[3] The instrument should be relevant because the more top universities there are within the state in which a given individual has received the university entrance qualification, the less costly it is for the student to go to such a university. This makes the additional assumption that the lack of top-universities does not alter the students' subject decision. The number of top universities should not impact the wage regression directly but only through the rankings (note that the wage equation controls in addition for federal states).

It seems likely that the instrument does not have an equal effect on each individual. Given that treatment effects are potentially heterogeneous, the ranking coefficient should represent the local average treatment effect, driven by the compliers of the instrument (Angrist & Pischke, 2008).

[3] I also included (relative) distances to the next top universities and neighboring states, but these instruments did not show an improvement of the first stage F-statistic and had no substantial effect on the coefficients.

4.5 Results

The regression results for the QS ranking are summarized in Table 4.2. In the first four columns, I present the OLS estimates and the IV estimates in the next four. The first column for both OLS and IV uses the general QS ranking. Columns (2) and (6) show the estimates for the dummy indicating whether the QS ranked a university for this area or not. Columns (3) and (7) use an indicator for the top quartile of universities per area of study of the QS ranking. As the 75th percentile's cut-off value is subjectively chosen, column (4) and (8) also show regression results for an indicator of the top decile.

Table 4.2 – Main OLS & IV regression results - QS ranking

	Dependent variable: Log monthly gross wage							
	OLS				IV			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
QS ranking	-0.0000				0.0010^{*}			
	(0.0002)				(0.0004)			
QS Indicator		-0.0053				0.0822^{*}		
		(0.0126)				(0.0322)		
Top QS quartile			0.0186				0.1036^{***}	
			(0.0119)				(0.0308)	
Top QS decile				0.0234				0.1337^{***}
				(0.0143)				(0.0338)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.526	0.526	0.526	0.526	0.522	0.522	0.522	0.521
Observations	6573	6573	6573	6573	6573	6573	6573	6573
F-Stat. 1. Stage					560	623	774	677

Note: The different columns include four different ranking variables. The first four columns are OLS estimates while the last four are IV estimates. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The QS ranking in column (1) has a coefficient of zero and is thus both economically and statistically insignificant. Simplifying the ranking into the QS dummy in column (2) even decreases the estimator but is still statistically insignificant. Next, since the literature has shown a non-linear effect of top-universities, in column (3), the ranking variable is replaced with a dummy for the top quartile of QS-ranked universities. Here, the coefficient increases to 0.0186, translating into an estimated monthly gross wage (henceforth referred to as "wages") increase by 1.86%. Though the effect would be of economic interest, it is not statistically significantly different from zero. Using a more selective definition of top universities, column (4) shows a 2.34% wage premium for the top decile, which is still statistically insignificant.

The coefficients for all four specifications increase when running the IV regressions. The QS ranking coefficient increases only marginally and is now equal to 0.0010 and statistically significant at the 5% level. The coefficient of the three dummies increases to 0.0822, 0.1036, and 0.1337, respectively, with the last two being highly statistically significant. The estimated wage increase of 8 to 13% is within the boundaries of effects found in the literature.

Though found in the literature in other countries as well, this effect seems high for a country with a low university hierarchy, which is why I further look at the RPA ranking in Table 4.3. Within this ranking, students - not a rating institution - decide what a top university is. Column (1) includes the mean high school GPA without converting it into a ranking. The coefficient is equal to -0.0384 and is significant at the 5% significance level. Thus, a better mean GPA by one (that means a lower GPA) is associated with a 3.84% increase in the gross wage for the observed graduates.[4] Using the RPA ranking leads to a coefficient equal to 0.0003 which is found to be statistically significant at the 5% level.[5] The small coefficient should be put in perspective: the ranking ranges from 0 to 100, and if a university improves according to this ranking definition, the increase in percentile would be more than one. Thus, a 10 percentage points improvement in the ranking is associated with a 0.3% increase in the gross monthly wages. To simplify the RPA ranking, the third column shows an estimate for a dummy for the top quartile of universities per area of study. Using this specification, I find an associated increase in wages of 1.32%, which is, however, insignificant. When comparing the bottom 90 to the top 10, the coefficient is equal to 0.0307, with a statistical significance at the 1% level. Therefore, graduating from a university of the top decile of the RPA ranking would predict a higher wage of 3%. To overcome the selection problem, I again use the number of top universities within the federal state of the higher education entrance qualification.

[4] In Germany grades range from 1 = best to 5 = worst.

[5] Results are similar if the ranking is based only on those with a GPA higher than 2.0 or on the top 10% high school students (per area of study).

As for the QS ranking, the coefficients increase and are all statistically significant. The wage premium for the top quartile is equal to 4.85% and for the top decile 8.25%. Thus, the RPA ranking suggests a smaller but still substantial wage premium.

Table 4.3 – Main OLS and IV regression results - RPA ranking

	Dependent variable: Log monthly gross wage							
	OLS				IV			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean HS GPA	−0.0384* (0.0164)				−0.1135* (0.0559)			
RPA ranking		0.0003* (0.0001)				0.0010* (0.0005)		
Top RPA quartile			0.0132 (0.0090)				0.0485* (0.0239)	
Top RPA decile				0.0307** (0.0114)				0.0825*** (0.0248)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.556	0.556	0.556	0.556	0.555	0.555	0.555	0.555
F-Stat. 1. Stage					728	858	1400	1600
Observations	10218	10218	10218	10218	10218	10218	10218	10218

Note: The different columns include four different ranking variables. The first four columns are OLS estimates while the last four are IV estimates. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Although one might expect lower IV estimates compared to OLS in the returns to education, this is not uncommon in the literature. Higher IV coefficients are reported in other papers as well, for instance in Angrist and Krueger (1991), Kling (2001), Lemke and Rischall (2003), Dee (2004), Milligan et al. (2004), Oreopoulos (2007), Kemptner et al. (2011), and Stephens and Yang (2014). Card (1995), for example, reports an OLS estimate of 8% and an IV estimate of 10 to 14%, which is a similar difference I find for the RPA ranking. However, the reasoning for this might be specific for Germany, given that papers mentioned above are mostly for the US. One usually suspects an ability bias, which leads to an upward bias of the OLS results. In the regression, I included the high school GPA and the federal state in which the individuals obtained the HEEQ as ability measures, which makes a large ability bias unlikely. Further, I included selection decisions of individuals for the type of high schools, university degrees as well as the type of university. If individuals opted, for example, for applied universities (similar for a HEEQ from vocational high schools), they could reveal less ambition or more interest in a practical degree, which helps to reduce an ability or motivation bias. Lastly, the (work) experience before graduation further proxies ambition, time preference, interest, and some signal-behavior of individuals. The downward bias of the OLS estimation might be due to the flat hierarchy. Lower-ability students could get to a good university because higher-ability students might not be willing to move far away "just" for a better university. A reason could be the desire to stay close to the family or budget constraints. Thus, the IV-coefficient is higher than the OLS-coefficient because students only end up at a good university because it is nearby. In other words, because the hierarchy of universities in Germany is rather flat, higher-ability students might also be concerned about where to live and face possible budget constraints of moving far. The LATE then identifies the compliers, who go to a top university just if it is close.

To check for the sensitivity, Tables C3 and C4 present estimation results dropping smaller areas of study, and the estimation is robust. The estimate changes only for the top decile of the RPA ranking when less than half of the observations are left in the sample. Dropping smaller universities, as shown in Tables C5 and C6, does not lead to substantially different estimation results either.

Given that both cut-offs of the top quartile and top decile are somewhat arbitrary, Figures C1 and C2 show the dummy coefficient with a rolling cut-off and its 90% confidence interval for the IV specification. Only for the RPA, when the top percentile is above 93, the coefficient drops and is insignificant. Thus, the results are overall very robust to different model specifications.

4.5.1 Wave specific regressions

The dataset includes wage information of two waves: one year after graduation and five years after graduation. The question naturally arises if the estimated wage premium is carried on from wave one to wave two or whether it is just found in one of them as in Weinstein (2017). Tables C7 and C8 show the wave specific estimation results for the IV regressions. This wave separation suggests an interesting pattern: for the QS ranking, the four different measures are only significant for the first wave (column 1 to 4) but not for the second (column 5 to 8). This difference between the waves would suggest that graduates from a top university benefit in terms of a higher starting salary, which equalizes at some later point of the career. This could suggest that the QS wage premium is a signal effect. Employers know the QS ranking and thus give these graduates higher wages. Graduates from other universities, however, catch up after they proved themselves five years after graduation.

For the RPA ranking, the case differs. Here, the coefficient for the top decile is significant in both waves, and the point estimate of the wage premia amount to 8.33% (wave 1) and 8.93% (wave 2), so they are fairly similar. This stable estimation suggests that students graduating from the top decile of the RPA ranking experience a wage premium early on and keep that advantage. For the top quartile, only the second wave is statistically significant and with a premium of 6.63% more prominent than the first wave of 2.51. This could mean that, after some time, graduates get the wage premium because they acquired more human capital at better universities, which is not known to the employer right away. However, graduates from the top decile of the RPA ranking might combine both the signal and the human capital acquisition.

4.5.2 Gender specific regressions

Labor participation and aspiration (among others) usually lead to different wage distributions of males and females. Thus, Tables C9 and C10 show IV regression results for the QS and the RPA rankings for men and women. For the QS ranking, the coefficients for men in columns (1) to (4) are all positive around 3 to 4% but not statistically significant. The coefficients are also smaller in magnitude compared to the gender-unspecific regressions in Table 4.2. These estimates are around the lower bound of the 99% confidence interval of the regressions in Table 4.2.

Consistent with the lowered ranking coefficients for males, the coefficients for the female regression specifications in columns (5) to (8) are larger and statistically significant.

Female graduates with a degree from the top decile experience a monthly gross wage premium of 12.42% up to 22.73%. One driver of this high impact might be the missing subject areas in the QS ranking.

For the RPA ranking, the case is less extreme but similar. The top decile leads to a statistically significant wage premium of 5.45% for men; the premium for women is equal to 9.20% and significant at the 1% level. Thus, the top university wage premium of the RPA ranking increases only by around one percentage point from 8.25 up to 9.20%.

In general, the wage premium for females is not necessarily surprising. Women face a gender wage gap due to several reasons. One reason is lowered competition seeking. Graduating from more competitive universities could be interpreted as a signal of ability and ambition by the employer. Or characterize these females by the higher ambition.

4.5.3 Subject-specific regressions

Next, I run subject specific regressions for the five most popular subject groups: languages, social sciences, math and natural sciences, medicine, and engineering. The results are shown in Tables C11 and C12. Here, one can see that both rankings lead to a wage premium in the subject group social sciences. Graduating from a QS-declared top university further gives a wage premium for medical students, while the RPA shows only an additional effect in the subject group engineering.

4.5.4 Comparing the QS and RPA rankings

Lastly, I want to compare the QS and RPA rankings in Table 4.4. To minimize the amount of output, I only show the IV regression results for the top decile of both rankings. Since the RPA ranking can be calculated for almost all students but the QS ranking only exists for students in certain fields of study, the number of observations differs. Thus, before including both rankings into one regression, I analyze the impact of the RPA ranking for the areas of study of the QS subsample. The estimated wage premium decreases from 8.25% for the full sample to 7.55%, significant at the 5% level. Thus, the main effect is not changed much. Then, column (2) looks at the female QS subsample because earlier results suggested that especially females benefitted. The effect increases slightly from 9.20% up to 10.62%, significant at the 10% level. Here as well, I find robust estimation results. The next two columns look at the first and second waves separately. For the RPA subsample, the effects amount to 8.33 and 8.93 for the respective waves. For the QS subsample, the effect for the first wave decreases to 5.36 and is not statistically significant anymore, while the effect for the second wave increases up to 10.68% and is still significant at the 5% level. This suggests that the benefit of a degree from a high ranked university does need time to show in the wage premium.

Table 4.4 – Regression results - QS & RPA ranking

	Dependent variable: Log monthly gross wage							
	RPA				RPA & QS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subsamples:	QS	Females	1st Wave	2nd Wave	QS	Females	1st Wave	2nd Wave
Top RPA decile	0.0755* (0.0370)	0.1062^{+} (0.0611)	0.0536 (0.0496)	0.1068* (0.0508)	0.0456 (0.0367)	0.0710 (0.0616)	0.0155 (0.0499)	0.0982* (0.0496)
Top QS decile					0.1301*** (0.0322)	0.2179*** (0.0517)	0.1659*** (0.0418)	0.0383 (0.0464)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.520	0.515	0.493	0.301	0.515	0.504	0.485	0.299
F-Stat 1.Stage	703	286	396	294	358	145	203	147
Observations	6574	3201	4113	2461	6574	3201	4113	2461

Note: The different columns include four different ranking variables. The first four columns are OLS estimates while the last four are IV estimates. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Columns (5) to (8) of the same table then include the top decile of the QS ranking to see which ranking matters more. One can see that for the general QS subsample, the females and the first wave, the QS ranking is highly significant while the RPA ranking is insignificant. The QS ranking is only statistically insignificant for the second wave, while the RPA ranking is statistically significant for this wave. This underlines that the more easily accessible ranking (QS), which is published by the QS, reported by newspapers and the universities, leads to a short-term wage premium, while the less accessible ranking (RPA) is more beneficial in the medium-term.

This could mean that the QS ranking serves as a signal for human resources departments but that not all factors of the evaluation of universities by the QS are relevant for graduates' success at a company. More specifically, it seems doubtful that academic specific measures, such as academic reputation, citations per faculty or international faculty ratio impact students' human capital acquisition in a relevant way for the occupations after graduation. On the other hand, the RPA ranking only uses the high school GPA, which could measure ability, motivation for learning, and initial endowment. The competition at these universities could be higher, leading to higher student achievements. The mean high school GPA is not broadcasted by the newspapers or universities, which is why there is no signaling and employers experience the value of the graduates within the first years of employment.

4.6 Conclusion

The literature shows a wage premium for graduates from elite universities, especially for the US, England, and Australia. I analyze the same question for Germany, a country with a relatively flat hierarchy of universities. Therefore, I use the graduate panel of the DZHW with graduates one and five years after finishing university.

To identify universities as "better" than others, I use two different approaches. The first approach is the QS ranking, which relies on typical university quality measures. The institute publishes these rankings yearly by fields of study since 2014 and used for promotion by (high ranked) universities. The second approach is a revealed preferences and acceptance ranking based on the mean high school GPA of students per cohort, university, and field of study. This ranking represent solely the general ability of the peers and partly the level of competition students face. Even though the hierarchy of German Universities is rather flat, I find a robust significant positive effect using both the QS and the RPA ranking on wages. Being in the top decile of the QS ranking, for example, gives a wage premium of around 13%. For the RPA ranking, the wage premium is around 8%. These wage premia are well in line with results of the literature.

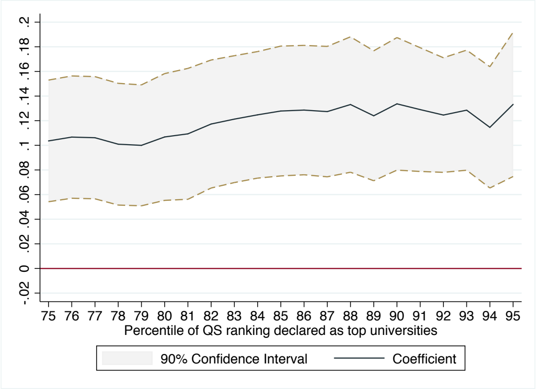
Aside from the small difference, one striking difference between the two rankings is revealed when regressing the first and second waves separately. Then, the QS ranking gives a wage premium only in the first wave. The RPA ranking gives a wage premium in both waves.

Moreover, the main profiteers appear to be females compared to males. This is in line with Belman and Heywood (1991) and Walker and Zhu (2011) who showed that women tend to benefit more from tertiary education compared to males. In this chapter, the main difference is that women benefit on the internal and not external margin.

Appendix

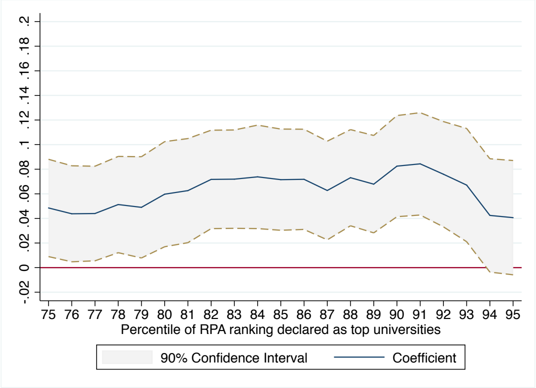
C.1 Figures

Figure C1 – Rolling cut-off value for indicator of top universities - QS ranking



Note: The graphs shows the regressions results for the IV regressions of the top percentile of the QS ranking, starting from the top 25th to the top 5th percentile. The anthracite solid line shows the coefficient and the gold dashed line shows border of the 90% confidence interval. The red solid line emphazises the the x-axis.

Figure C2 – Rolling cut-off value for indicator of top universities - RPA ranking



Note: The graphs shows the regressions results for the IV regressions of the top percentile of the QS ranking, starting from the top 25th to the top 5th percentile. The anthracite solid line shows the coefficient and the gold dashed line shows border of the 90% confidence interval. The red solid line emphazises the the x-axis.

C.2 Tables

Table C1 – Summary statistics of areas of study and federal states

	Obs	Mean	SD	Mean	SD	Mean	SD
	General sample			RPA subsample		QS subsample	
Area of Study							
Linguisitcs and cultural sciences	16453	0.01	0.10	0.01	0.11	0.02	0.13
Protestant theology	16453	0.00	0.06	0.00	0.06	0.00	0.00
Catholic theology	16453	0.01	0.08	0.01	0.08	0.00	0.00
Philosophy	16453	0.00	0.05	0.00	0.05	0.00	0.00
History	16453	0.01	0.10	0.01	0.11	0.00	0.00
Library science	16453	0.01	0.12	0.02	0.14	0.00	0.00
General and comparative literature studies	16453	0.01	0.07	0.01	0.07	0.00	0.00
Classical philosophy	16453	0.00	0.02	0.00	0.02	0.00	0.00
German studies	16453	0.04	0.19	0.04	0.19	0.00	0.00
Anglistics	16453	0.02	0.13	0.02	0.13	0.03	0.17
Romanistics	16453	0.01	0.08	0.01	0.08	0.00	0.00
Slavistics	16453	0.00	0.01	0.00	0.01	0.00	0.00
Non-European linguistics and cultural sciences	16453	0.00	0.02	0.00	0.02	0.00	0.00
Cultural studies in a wider sense	16453	0.00	0.04	0.00	0.04	0.00	0.00
Psychology	16453	0.02	0.12	0.02	0.13	0.03	0.16
Educational sciences	16453	0.04	0.21	0.04	0.20	0.04	0.20
Sports	16453	0.01	0.08	0.01	0.08	0.00	0.00
Business and social studies, generally	16453	0.01	0.08	0.01	0.08	0.00	0.00
Regional sciences	16453	0.00	0.01	0.00	0.01	0.00	0.00
Political sciences	16453	0.01	0.10	0.01	0.10	0.02	0.13
Social sciences	16453	0.01	0.11	0.01	0.12	0.02	0.14
Social services	16453	0.05	0.21	0.05	0.21	0.00	0.00
Legal studies	16453	0.04	0.19	0.01	0.08	0.01	0.10
Administrative sciences	16453	0.01	0.08	0.01	0.08	0.00	0.00
Economic sciences	16453	0.15	0.36	0.17	0.37	0.26	0.44
Industrial engineering	16453	0.04	0.20	0.04	0.20	0.00	0.00
Mathematics	16453	0.03	0.16	0.03	0.17	0.04	0.21
Computer Sciences	16453	0.05	0.23	0.06	0.23	0.09	0.28
Physics, astronomy	16453	0.01	0.09	0.01	0.09	0.01	0.11
Chemistry	16453	0.01	0.11	0.01	0.11	0.02	0.13
Pharmaceutics	16453	0.00	0.02	0.00	0.02	0.00	0.02
Biology	16453	0.02	0.14	0.02	0.13	0.03	0.16
Geosciences	16453	0.00	0.06	0.00	0.06	0.00	0.00
Geography	16453	0.01	0.11	0.01	0.12	0.02	0.15
Health sciences, generally	16453	0.01	0.10	0.01	0.11	0.00	0.00
Human medicine	16453	0.05	0.23	0.06	0.23	0.09	0.28

Veterinary medicine	16453	0.01	0.11	0.01	0.11	0.00	0.00
Landscape management	16453	0.02	0.12	0.02	0.13	0.00	0.00
Agricultural sciences	16453	0.02	0.13	0.02	0.13	0.00	0.00
Forestry, wood industry	16453	0.01	0.07	0.00	0.07	0.00	0.00
Food sciences and home economics	16453	0.01	0.08	0.01	0.08	0.00	0.00
Engineering, generally	16453	0.00	0.07	0.01	0.07	0.00	0.00
Mining, metallurgy	16453	0.00	0.03	0.00	0.03	0.00	0.04
Mechanical engineering, process engine	16453	0.08	0.28	0.08	0.28	0.13	0.33
Electrical engineering	16453	0.03	0.18	0.04	0.18	0.05	0.23
Traffic engineering, nautical science	16453	0.01	0.12	0.01	0.12	0.00	0.00
Architecture, interior design	16453	0.03	0.17	0.03	0.18	0.05	0.22
Spatial planning	16453	0.00	0.06	0.00	0.05	0.00	0.00
Civil engineering	16453	0.03	0.17	0.03	0.17	0.05	0.21
Surveying	16453	0.01	0.11	0.01	0.11	0.00	0.00
Art, aesthetics, generally	16453	0.00	0.06	0.00	0.06	0.00	0.00
Fine art	16453	0.00	0.03	0.00	0.03	0.00	0.00
Design	16453	0.01	0.11	0.01	0.11	0.00	0.00
Performing art, film and television	16453	0.00	0.04	0.00	0.03	0.00	0.00
Music	16453	0.00	0.07	0.00	0.07	0.00	0.00
Federal states							
Schleswig-Holstein	16252	0.03	0.17	0.03	0.17	0.03	0.17
Hamburg	16252	0.02	0.14	0.02	0.14	0.02	0.14
Niedersachsen	16252	0.10	0.30	0.10	0.31	0.10	0.31
Bremen	16252	0.01	0.08	0.01	0.08	0.01	0.09
Nordrhein-Westfalen	16252	0.16	0.37	0.17	0.38	0.16	0.37
Hessen	16252	0.06	0.24	0.06	0.24	0.06	0.24
Rheinland-Pfalz	16252	0.04	0.20	0.04	0.19	0.04	0.20
Baden-Württemberg	16252	0.15	0.36	0.14	0.35	0.14	0.35
Bayern	16252	0.15	0.36	0.13	0.33	0.14	0.35
Saarland	16252	0.01	0.08	0.01	0.08	0.01	0.07
Berlin	16252	0.03	0.17	0.04	0.18	0.04	0.19
Brandenburg	16252	0.04	0.18	0.04	0.19	0.04	0.19
Mecklenburg-Vorpommern	16252	0.02	0.15	0.02	0.15	0.02	0.15
Sachsen	16252	0.09	0.29	0.10	0.30	0.09	0.29
Sachsen-Anhalt	16252	0.04	0.19	0.04	0.19	0.04	0.19

Note: The table adds the summary statistics for areas of study as well as the federal states in which the graduates obtained their higher education entrance qualification, pooled over both survey waves. The other variables are presented in Table 4.1. The first column of observations only excludes individuals without a full-time job. The second column of observations contains only individuals for which all information in the RPA regression are available. The third further drops individuals with areas not ranked by the QS. The RPA-subsample includes 10218 individuals whereas the QS-subsample includes only 6573 individuals. *Source*: DZHW Graduate Panel 2005 and 2009, own calculations.

Table C2 – QS ranking

Subject of study	Area of study	Universities ranked per year			
		2014	2015	2016	2017
Language and cultural sciences					
	English studies	200	300	300	300
	Media studies	200	200	200	200
Sports					
	Sport studies	0	0	0	100
Legal, economic and social sciences					
	Business administration	0	200	200	300
	Education science	200	200	300	300
	Law	200	200	200	300
	Politics and Sociology	200	200	200	300
	Psychology	200	200	200	300
	Economics	200	200	300	400
	Public management and governance	0	0	100	100
Mathematics, natural sciences					
	Biology	200	400	500	500
	Chemistry	200	200	200	300
	Geography	200	200	200	200
	Computer sciences	200	400	500	500
	Mathematics	200	400	400	400
	Pharmacy	200	200	200	300
	Physics	200	500	400	500
Medicine, health care sciences					
	Medicine	200	400	500	500
Veterinary medicine					
	No area of study from this subject is included in the QS ranking				
Agricultural forestry and nutritional sciences					
	No area of study from this subject is included in the QS ranking				
Engineering sciences					
	Architecture	0	100	100	200
	Building and environmental engineering	200	200	200	200
	Electrical engineering	200	300	400	400
	Engineering	200	300	300	400
Art, aesthetics					
	No area of study from this subject is included in the QS ranking				

Note: The table shows the areas of study which are ranked by the QS. There are no areas of the subject groups Veterinary medicine, agricultural forestry and nutritional sciences, and art and aesthetics. Sports is not included in the QS-analysis, because not a single German university was ranked.

Table C3 – Excluding small areas of study - QS ranking

	Dependent variable: Log monthly gross wage					
	Areas > 40	Areas > 50	Areas > 60	Areas > 70	Areas > 80	Areas > 90
Top QS decile	0.1437*** (0.0410)	0.1383** (0.0436)	0.1788*** (0.0493)	0.1847*** (0.0495)	0.1959*** (0.0512)	0.1187 (0.0770)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.536	0.491	0.473	0.470	0.408	0.382
F-Stat. 1. Stage	474	430	354	351	334	178
Observations	5317	4723	4019	3948	3683	2537

Note: The sample is restricted to individuals in an area of study for which we have more than a certain number of graduates, specified in top of the column. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C4 – Excluding small areas of study - RPA ranking

	Dependent variable: Log monthly gross wage					
	Areas > 40	Areas > 50	Areas > 60	Areas > 70	Areas > 80	Areas > 90
Top RPA decile	0.1259*** (0.0315)	0.1323*** (0.0325)	0.1462*** (0.0356)	0.1556*** (0.0422)	0.1044^{+} (0.0588)	−0.0181 (0.1039)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.596	0.585	0.589	0.542	0.425	0.391
F-Stat. 1. Stage	608	578	500	376	214	81
Observations	6772	6178	5474	4585	3683	2537

Note: The sample is restricted to individuals in an area of study for which we have more than a certain number of graduates, specified in top of the column. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C5 – Excluding small universities - QS ranking

	Dependent variable: Log monthly gross wage					
	Unis > 400	Unis > 500	Unis > 600	Unis > 700	Unis > 800	Unis > 900
Top QS decile	0.1610*** (0.0330)	0.1593*** (0.0330)	0.1644*** (0.0328)	0.1662*** (0.0328)	0.1656*** (0.0331)	0.1577*** (0.0322)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.514	0.515	0.517	0.516	0.516	0.518
F-Stat. 1. Stage	705	706	717	716	710	739
Observations	6185	6026	5828	5719	5630	5507

Note: The sample is restricted to individuals from universities for which we have more than a certain number of graduates, specified in top of the column. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C6 – Excluding small universities - RPA ranking

	Dependent variable: Log monthly gross wage					
	Unis > 400	Unis > 500	Unis > 600	Unis > 700	Unis > 800	Unis > 900
Top RPA decile	0.0832** (0.0268)	0.0796** (0.0263)	0.0842** (0.0273)	0.0868** (0.0276)	0.0812** (0.0280)	0.0848** (0.0284)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.550	0.550	0.552	0.554	0.553	0.555
F-Stat. 1. Stage	998	1100	1046	1027	1001	980
Observations	9602	9323	9033	8846	8694	8470

Note: The sample is restricted to individuals from universities for which we have more than a certain number of graduates, specified in top of the column. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C7 – Wave specific IV regression results - QS ranking

	Dependent variable: Log monthly gross wage							
	First wave				Second wave			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
QS ranking	0.0012* (0.0005)				0.0007 (0.0005)			
QS Indicator		0.0975* (0.0384)				0.0619 (0.0457)		
Top QS quartile			0.1446*** (0.0371)				0.0424 (0.0425)	
Top QS decile				0.1670*** (0.0418)				0.0518 (0.0471)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.488	0.488	0.486	0.485	0.298	0.297	0.300	0.301
F-Stat. 1. Stage	466	514	680	610	320	362	412	360
Observations	4112	4112	4112	4112	2461	2461	2461	2461

Note: The different columns include four different ranking variables. The first wave is about one year after graduation while the second wave is four to five years after graduation. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C8 – Wave specific IV regression results - RPA ranking

	Dependent variable: Log monthly gross wage							
	First wave				Second wave			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean HS GPA	−0.0593 (0.0722)				−0.1537* (0.0756)			
RPA ranking		0.0005 (0.0006)				0.0014* (0.0007)		
Top RPA quartile			0.0251 (0.0305)				0.0663* (0.0327)	
Top RPA decile				0.0833* (0.0327)				0.0893** (0.0338)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.537	0.537	0.537	0.537	0.349	0.347	0.346	0.347
F-Stat 1.Stage	418	494	795	901	296	343	580	649
Observations	6408	6408	6408	6408	3810	3810	3810	3810

Note: The different columns include four different ranking variables. The first wave is about one year after graduation while the second wave is four to five years after graduation. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C9 – Gender specific IV regression results - QS ranking

	Dependent variable: Log monthly gross wage							
	Man				Women			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
QS ranking	0.0003 (0.0005)				0.0015** (0.0006)			
QS Indicator		0.0284 (0.0445)				0.1242** (0.0475)		
Top QS quartile			0.0281 (0.0418)				0.1715*** (0.0470)	
Top QS decile				0.0417 (0.0442)				0.2273*** (0.0545)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.471	0.471	0.471	0.471	0.510	0.510	0.509	0.505
F-Stat. 1.Stage	318	337	445	392	255	293	318	286
Observations	3372	3372	3372	3372	3201	3201	3201	3201

Note: The different columns include four different ranking variables. The sample is separated by gender. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C10 – Gender specific IV regression results - RPA ranking

	Dependent variable: Log monthly gross wage							
	Men				Women			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean HS GPA	−0.0713 (0.0882)				−0.1139 (0.0789)			
RPA ranking		0.0007 (0.0008)				0.0010 (0.0007)		
Top RPA quartile			0.0314 (0.0388)				0.0480 (0.0331)	
Top RPA decile				0.0545 (0.0410)				0.0920** (0.0356)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.499	0.498	0.498	0.498	0.537	0.537	0.538	0.537
F-Stat 1. Stage	189	204	387	410	304	387	525	584
Observations	4706	4706	4706	4706	5512	5512	5512	5512

Note: The different columns include four different ranking variables. The sample is separated by gender. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses. Top RPA quartile refers to the top quartile of the RPA ranking.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C11 – Specific subject groups - QS ranking

	Dependent variable: Log monthly gross wage				
	Languages	Social Sciences	Math. & Nat. Sc.	Medicine	Engineering
Top QS decile	−0.0009 (0.0009)	0.0027* (0.0012)	0.0011 (0.0007)	0.0017* (0.0007)	−0.0010 (0.0008)
Control variables	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.406	0.367	0.612	0.490	0.483
First state F-Statistics	103	121	96	137	137
Observations	734	2012	1402	562	1836

Note: The sample is restricted to certain subject areas of study. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C12 – Specific subject groups - RPA ranking

	Dependent variable: Log monthly gross wage				
	Languages	Social Sciences	Math. & Nat. Sc.	Medicine	Engineering
Top RPA decile	0.0012 (0.0010)	0.0015^{+} (0.0008)	0.0002 (0.0021)	-0.0009 (0.0009)	0.0031^{*} (0.0015)
Control variables	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.516	0.457	0.602	0.586	0.478
First state F-Statistics	195	271	44	175	43
Observations	1846	3068	1448	791	2209

Note: The sample is restricted to certain subject areas of study. Only the IV estimates are presented in the table. Individual cluster and heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 5

Practice makes perfect? Self-testing with external rewards[*]

[*] This chapter is based on: Schwerter, J., J. Bleher, T. Dimpfl and K. Murayama (2020): Practice makes perfect? Self-testing with external rewards, unpublished manuscript, University of Tübingen and University of Reading.

5.1 Introduction

During the last ten years, university teaching saw a significant increase in the use of online tools, and another boost of online lectures in the first half of 2020 due to the COVID-19 pandemic leading to a closure of lecture halls and a switch to online teaching. The literature evaluating this new way of academic teaching has been growing accordingly (for example, Broadbent & Poon, 2015; B. W. Brown & Liedholm, 2002; M. G. Brown, 2016; Butler, 2010; D. Coates et al., 2004; Figlio et al., 2013; Fischer, Zhou et al., 2019; Kizilcec, Pérez-Sanagustín & Maldonado, 2017; O'Flaherty & Phillips, 2015; Paechter, Maier & Macher, 2010; Thai et al., 2017; D. Xu & Jaggars, 2014). A lot of these studies, however, focus on differences between face-to-face lectures and online lectures. Little is known about practicing the study materials online.

Therefore, this chapter studies the question of whether extra online practice is beneficial for the students in terms of better final grades. To shed some light on this issue, we investigate a scenario of rewarded and voluntary practice during the semester. This chapter uses an observational study with students from the University of Tübingen course *Mathematics for Economics and Business Administration* for first-semester students. Within this course, students were allowed to take three midterm tests to achieve (at best) six additional points for the final grade (which is calculated out of 60 points in total). The participation in the midterms was voluntary. Then, we allowed the students to retake each midterm as often as they wanted without the additional chance to earn extra points. We then analyze if students who took part in the midterms and made use of the additional practice opportunity achieve a higher number of points in the final exam (without including the reward of the midterms). In addition to the midterms, students could voluntarily practice, collect scores for correct solutions and be ranked in a web app called 'a matrix a day' to improve their skills in linear algebra. To identify the effect of participation in the midterms, additional practice, and the matrix-app, we control for student specific features and collect a set of performance and personality measures. Since participation and performance are likely to be driven by motivation, ability, or personality traits, we surveyed, among others, achievement goals, items of the expectancy-value theory, present bias preferences and the big five personality traits.

Controlling for these personal characteristics should eventually lead to a model which allows to identify a causal practice effect. We were not allowed to treat students differently, i.e., randomly assigning students to a treatment group who is allowed to participate in the additional practice opportunities offered, and a control group who is not. With the help of the before-mentioned control variables, we are able to control for the most important drivers of student achievement. Our results indicate that participation in the midterms and subsequent voluntary practice of them increase the final number of grade points between 2.6 and 5, depending on the applied methodology. The performance in the midterms has a positive impact on the exam grades, similar to the performance in the matrix-app.

Using a variety of variable selection methods (Lasso, Random Forest, and xgBoost), we also look for important predictors among our control variables for the exam points, the practice participation and outcome variables. We are thereby, to the best of our knowledge, the first to compare such a rich set of (psychological) measures to predict university exam grades.

The article proceeds as follows. Section 5.2 reviews the literature before we describe the specific setting on the chapter in Section 5.3. Then, Section 5.4 describes the data and Section 5.5 the econometric model. Section 5.6 present the results and Section 5.7 concludes.

5.2 Literature

This chapter is related to three strands of the education literature. First, we seek to explain learning success in terms of exam grades and this chapter is therefore related to the literature which deals with exam grade prediction. Second, our setting is within a class of mathematics and we, therefore, relate to studies considering math and statistics courses in particular. Finally, due to the online nature, our results also relate to studies concerning e-learning and self-testing. These aspects are laid out in more detail below.

Still, there are further related articles worth mentioning. For example, Broadbent and Poon (2015) highlight that peer learning is a very important aspect in online teaching which we approach, to some extent, using the matrix app to build an at least perceived group environment. M. G. Brown (2016) reviews the empirical literature which considers face-to-face teaching augmented with online tools. This is the setting of this chapter as well and we take up on his suggestion to explore methods in higher education further. Lastly, Thai et al. (2017) show that a blended learning design which we also have in the context of the course *Mathematics for Economics and Business Administration* is beneficial for learners. In particular, they find a significant effect on self-efficacy and intrinsic motivation.

This chapter contributes to the rich literature on the prediction of exam grades. McKenzie and Schweitzer (2001) show that for first-year Australian University students, previous academic performance, integration into the university, self-efficacy, and employment responsibility are essential predictors for exam grades. For Austria, Paechter et al. (2010) find that achievement goals and students' motivation are of high importance for their success. This is supported by findings of Komarraju and Nadler (2013) who, in addition, identify effort regulation and help-seeking behavior as important outcome predictors. Bailey and Phillips (2016) confirm the high positive impact of intrinsic motivation for first-year students. Rimfeld, Kovas, Dale and Plomin (2016), among others, emphasize the importance of personality traits such as the big five, and that grit, i.e., perseverance and passion for long-term goals, is only little related to exam achievements. Honicke and Broadbent (2016) provide a more comprehensive review. For the purpose of this chapter, we rely on these predictive variables to control for important confounders of students' achievement. This chapters' novel contribution to the existing literature is to compare several of those named measures and identify the most important ones.

This chapter is also related to a strand of the literature focussing on anxiety and more general performance problems in statistics and mathematics courses at the university level. Finney and Schraw (2003) show that current statistics self-efficacy and general self-efficacy to learn statistics of undergraduate students in the educational psychology department are correlated with performance in college statistics courses. Lane, Hall and Lane (2004) confirm the link between self-efficacy and statistics performance among sports degree seeking students. They use a voluntary hand-in worksheet in the middle of the semester to measure the progress of students. This chapter is related to these results due to the self-concept items in the expectancy-value theory (EVT) used as control variables and due to our online practice modules. Macher, Papousek, Ruggeri and Paechter (2015) show that statistics anxiety may help at the beginning of the semester to start studying, but is correlated with worse grades in the final exam. In the EVT, we have measures for the opportunity cost, effort, and emotions. The latter should relate to the anxiety with respect to statistics, an important predictor for missing practice and a bad exam grade. Acee and Weinstein (2010) find that value-reappraisal intervention increases exam performance, but only for certain instructor subgroups.

Finally, we do not only survey psychology measures and see if those explain exam grades. We also add to the e-learning literature by including the voluntary online-exercises with which students can practice during the semester. Thus, we add to the self-testing literature. Rodriguez, Fischer, Zhou, Warschauer and Massimelli (2016), Rodriguez, Kataoka et al. (2018), and Rodriguez, Rivas, Matsumura, Warschauer and Sato (2018) underline the importance of spacing and self-testing in learning and that they improve achievement outcomes of students.

H. Park, Behrman and Choi (2018) confirm the results on spacing. Using clickstream data, procrastinators perform significantly worse in exams. This group benefits from higher levels of regularity, while it is the other way around for non-procrastinators. Baker, Evans, Li and Cung (2019) aimed to improve student's time management by letting students scheduling their online lectures. Students who were allowed to schedule it on their own were better in the first quiz, especially those with low reported time management skills. This effect, however, did not persist over time and did not lead to better exam grades. The intervention did not show any effect on students' behavior in terms of cramming, procrastination, or the time at which students studied. Fischer, Zhou et al. (2019) presents results for a three weeks online preparation course in chemistry. They show participation improved exam grades by one-third of a letter grade. Especially at-risk students benefitted from online preparation. Thus, our results add to the question of spacing, participating, and performance of self-testing (or practicing).

5.3 Description of the course and the practice environment

5.3.1 Course information

Mathematics for Economics and Business Administration is a compulsory module in the first semester of all bachelor programs in economics and business administration (major and minor) at the University of Tübingen. Every year, about 350 students take the exam in this course, while some more students are registered on the open-source learning management system of the university (who ultimately decide not to take the exam).[1] The course has three voluntary midterms that are conducted using the built-in test feature of ILIAS. The midterms were first and foremost designed such that students have an opportunity to test their knowledge during the semester. In order to set an incentive for participation, the midterms are rewarded with extra points which count towards the final exam grade. In their context, we are interested in two aspects. First, who actually takes the midterm exams and, second, how much the students benefit from participating in terms of the exam grade at the end (without the extra points due to the midterms).

To provide additional opportunities for practicing, we make the midterms available as

[1] The university uses the open-source online learning management system ILIAS.

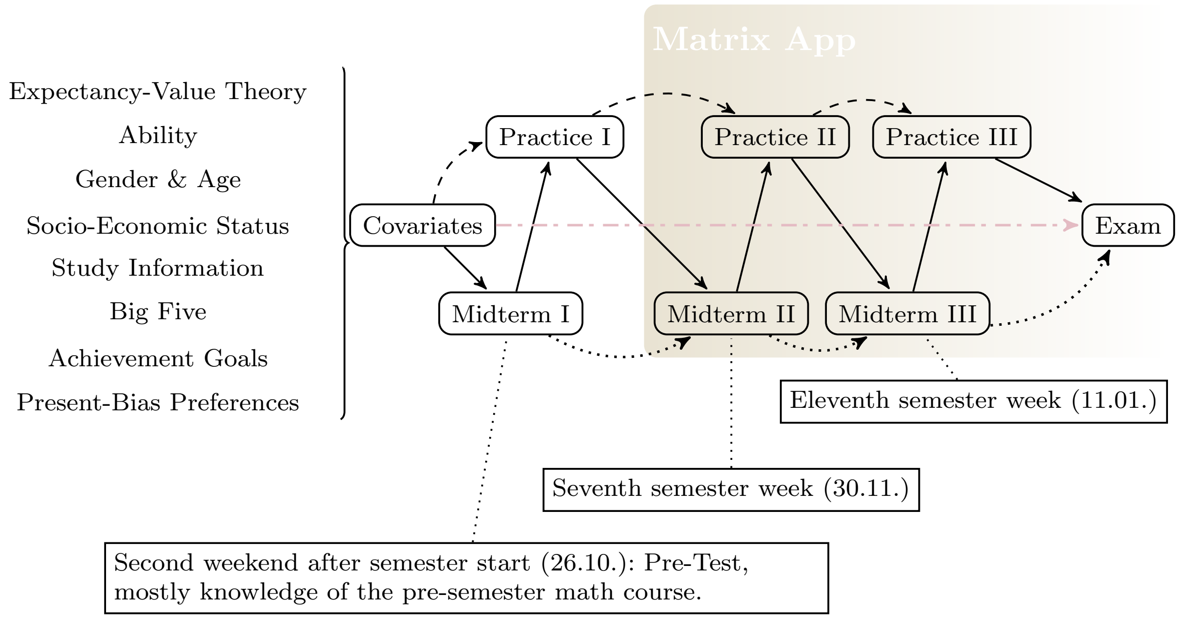
an online exercise tool with the same problems but changing numbers after running for the first time to gain extra credit. Students can then use them solely for practice without additional external rewards. The e-learning exercises (midterm and its voluntary use after that) are an easily accessible tool for students studying during the semester to avoid procrastination. This should, if students participate, also help to keep track during the semester. In the best-case scenario, students can follow the course better and are not left behind. Therefore, it is interesting to see who uses the additional online exercises and how both groups, users and non-users, perform in the final exam.

An additional tool that covers content of the first half of the semester, is the matrix app (called “A matrix a day”, MAD). The idea is to give students the opportunity to repeat tasks that are covered in the first half of the semester, also during the second half. This is important, as the first part of the lecture always seems easy, while the second appears more complicated, irrespective of what is taught first or second (anecdotal evidence of the lecturer). The app generates a matrix. Students are asked to calculate certain features like the determinant, eigenvalues, and, if possible, the inverse. They may register and enter their result into the app and, depending on their solution, obtain points. They can also opt to participate in a ranking. Students may also set an email reminder so that they are reminded each day when a new matrix had been generated. The six best performing students on the app were rewarded with a 20 Euro shopping voucher at the end of the semester. Students were able to enter the competition at any point during the semester. They were, however, not able to submit solutions for matrices of past days. On the webpage of the app, we also provided the history of the past week so that, in particular, shortly before the exam, additional study material was available.

5.3.2 Design of online exercises

The conceptual design of the e-learning exercises which accompany the course *Mathematics for Economics and Business Administration* is graphically depicted in Figure 5.1 and illustrates all possible ways to go through the course to the final exam. For students who do not make use of any of the practice opportunities provided, the path illustrated by the red dashed-dotted line would be applicable. These students obtain their final grade without focusing on the midterms or the matrix app such that their grades ultimately depend on their own ability to follow the course, influenced by personal characteristics like socio-economic status, preferences, and goals.

Figure 5.1 – Study design



Note: The figure describes the design of the practice part in the course *Mathematics for Economics and Business Administration* at the University of Tübingen. The dates refer to the winter semester 2019/20.

By contrast, the solid line represents motivated students or at least students who are willing to put in the maximum effort. Those students take part in all the midterms and use the online exercises to prepare for the exam as well as possible. The dotted line at the bottom represents the path of students who participate in the midterms only without completing any additional e-learning exercises. It is also possible that students participate partially, skipping one or more midterms and practice opportunities. Those possibilities are not displayed in Figure 5.1 to keep the overview concise. The matrix app was made available on November 8, 2019, and announced on November 13, 2019 in class. Students were free to use it at leisure. The shading in Figure 5.1 indicates that most of them had a look at the app immediately when it was set online, but the frequency of use declined towards the exam.

To conduct the analysis, we needed to collect information about the students' personal characteristics in class. This was done with a survey at the beginning of the semester. The survey, together with the study, was announced during the first lecture, which took place on October 17, 2019. Students were then given time until the following Sunday (October 20) to complete the online questionnaires. To achieve a high participation rate, we made use of a rattle with three iPads and 75 shopping vouchers worth 20 Euro each (to be used in shops in Tübingen) as prizes. The likelihood to win at least one of the prizes was roughly 25%. The lottery was implemented live in class on October 25, 2020, and the iPads were handed over directly while the vouchers needed to be picked up later.[2]

Table 5.1 provides an overview of the participation rates in the different activities offered during the semester. 378 students took at least one of the midterm tests, while 107 made use of the opportunity to retake them. 108 students used the matrix app. Lastly, 336 students were registered for the final exam (E) on February 10, 2020, out of which 56 did not present themselves for the exam on the exam date.[3] Regarding the survey, we have 325 students who filled out at least one of the questionnaires.

Table 5.1 – Cardinality of intersection sets

$\cap$	S	Z_i	O	M	E
S	325				
Z_i	305	378			
O	88	385	107		
M	99	108	41	108	
E	267	318	99	108	336

Note: The table displays the cardinality of the intersection sets among the various groups. S denotes the number of students who took the survey, Z_i is the number of students who participated in at least one of the intermediate tests. O denotes the number of students who practiced using the intermediate tests again. M is the number of students who used the matrix app and E are those who took the final exam.

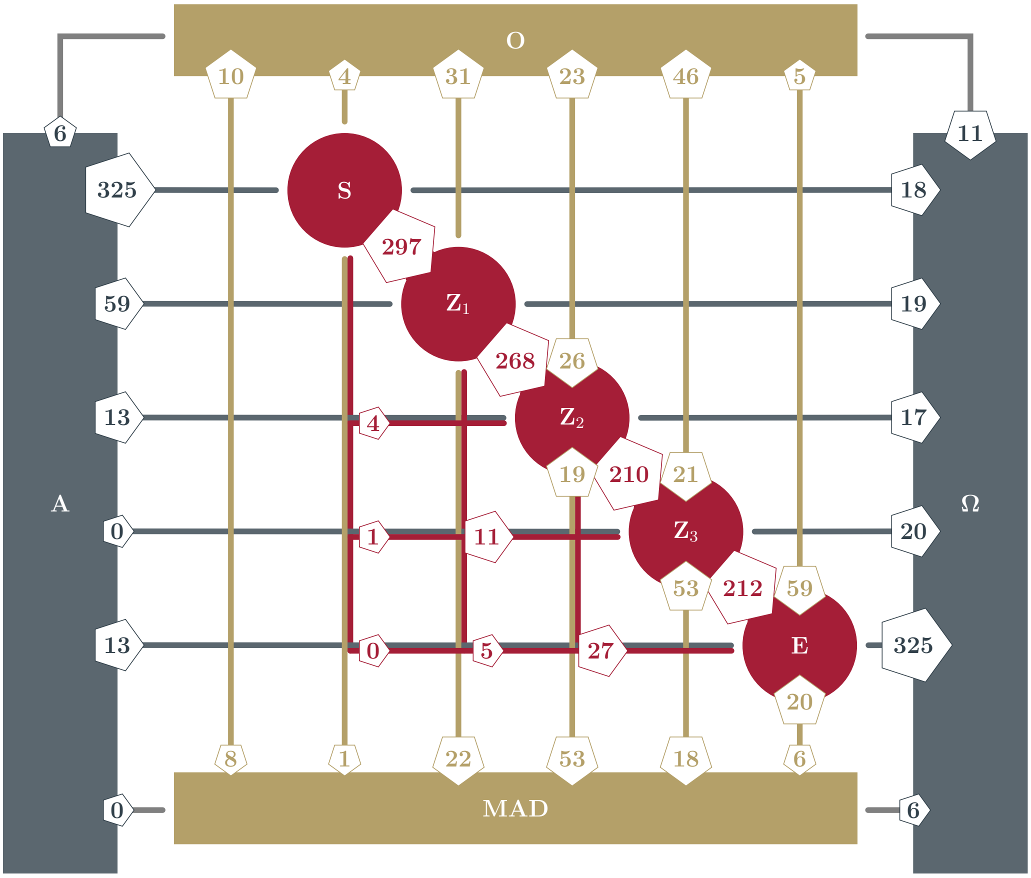
To provide more insights into the dynamics of participation and non-participation in

[2] Interestingly, there were six students who never collected their prizes.

[3] Comparing these numbers with past years does not show anything unusual.

the different activities, Figure 5.2 provides a schematic illustration of possible paths taken by students. As can be seen, all participants in the survey have also taken the first midterm test. However, participation declined as only subsets of students then took the later midterms. 33 students only took the first one and then proceeded directly to the exam, leaving possible additional exam points on the table. This is particularly interesting as the points also count towards passing the exam and might, therefore, be very valuable if only very few points are missing from the pass barrier (which is ex ante 45 points). As was already evident from Table 5.1, the participation in the voluntary, additional exercises (denoted O and MAD) became sparse towards the exam.

Figure 5.2 – Migration between states



Note: The figure describes the migration of students from one state to the next during the semester. A denotes all students that at least participated at one point during the semester. This could mean that they took the survey S, one of the three intermediate tests (Z_1, Z_2 ,Z_3), participated in practice opportunities O, participated in the matrix app MAD or wrote the exam E. All students have to eventually left the system in Ω. The polygons point towards the migration direction.

5.4 Data

Table 5.2 presents summary statistics for the exam, the midterms, the additional practice, the matrix app as well as all control variables. The latter include demographic information as well as psychological measures such as items of the expectancy-value theory, the big five, present bias preferences, and achievement goals. Additionally, we asked students about their subjective goals. There are 280 individuals who took the exam at the end of the semester.[4] Out of these, we have 175 students with full information, called complete cases sample from now on. Thus, for 105 students, we have at least one item missing. Out of the 105 students in the incomplete cases sample, there are about 50 students who provided at least some information into the survey. Some variables, however, like 'Mastery avoidance', have only 35 additional entries.

We constructed the practice variables in the following way: we added the number of midterms and extra practice (of the midterms) attempted by the students. The outcome variable then ranges from zero to six. Theoretically, it could go to infinity because students were allowed to repeat the practice offers as often as they wanted, but that did not happen. On the contrary, the maximum additional attempt was two. We then converted the points obtained in each trial to percentages of the respective midterm/practice and took the mean, considering only the number of midterms/practice tests a student really took. If students never participated, they were assigned zero points. For the matrix app, we have the number of submissions (as a measure for how often the app was used). For each submitted solution, students could obtain up to 11 points. In order to measure students' practice performance in the app, we take the mean percentages of the submitted solutions.

In order to avoid assumptions needed for the imputation of missing variables, we conduct

[4] There is a retake opportunity for the exam, which is excluded deliberately from this chapter due to the COVID-19 situation that followed after February 2020. In a normal semester, the exam takes place in April, before the summer term starts. In 2020, the exam was held in June, in the middle of the ongoing summer semester. Hence, the situations under which students took the exam are not comparable.

the analysis on complete cases only. This is valid if we can assume that the missing cases are approximately random. Comparing the descriptive statistics in Table 5.2 supports this assumption: the means of the full sample (with missing values), the complete case sample, and the incomplete-cases sample show only a few minor differences. Students who did not complete all questions have, on average, 6 points less in the exam, have practiced a bit less but made use of the MAD more often. It seems that more students who have a business or economics minor did not fully complete the survey. Among the students that did not complete the survey, 27 students attempted to take the exam before. They either did not succeed in past exams or did not show up due to sickness or other reasons. Among the students who completed the survey, there are only three students who repeated the exam. One of the repeaters took the survey but did not complete it. This suggests that students who have to repeat the exam, on the vast majority, did not show up in the first lecture when the survey was announced. Anecdotal evidence and teaching experience suggest that these students often feel that they already know the material from previous semester(s), so that they may skip the first few lectures. As information about whether somebody repeated the exam or not is available and can be controlled for, we can analyze a possible selection bias stemming from the omission of the repeaters. Thus, we defer further discussion to Section 5.6.

Table 5.2 – Descriptive statistics

	Full sample			Complete obs.			Incomplete obs.		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Exam (outcome)									
Grade points final exam	280	41.19	17.06	175	43.58	17.19	105	37.21	16.14
Practice (participation and performance)									
Num. midterm/practice taken	280	3.10	1.00	175	3.25	0.87	105	2.85	1.15
Points in midterms/practice	280	67.00	19.18	175	71.10	15.85	105	60.18	22.18
Submission MAD	280	2.74	8.38	175	26.65	36.92	105	21.55	37.21
Percentage MAD	280	24.74	37.05	175	2.53	7.91	105	3.10	9.14
Indiviual characteristics									
Female	280	0.56	0.50	175	0.56	0.50	105	0.55	0.50
High school GPA	226	2.08	0.60	175	2.07	0.59	51	2.12	0.66
Advanced math in HS	219	0.83	0.38	175	0.85	0.36	44	0.77	0.42
Last math grade in HS	226	2.62	1.10	175	2.59	1.09	51	2.71	1.14
International studies	280	0.41	0.49	175	0.44	0.50	105	0.35	0.48

Sports degree	280	0.08	0.26	175	0.05	0.22	105	0.11	0.32
Minor	280	0.16	0.37	175	0.13	0.33	105	0.22	0.42
Work to finance studying	210	0.22	0.41	175	0.19	0.40	35	0.34	0.48
Semester	225	1.23	1.10	175	1.26	1.20	50	1.12	0.63
Re-taking exam	225	2.01	0.21	175	2.00	0.19	50	2.04	0.28
Expectancy value theory									
Self-concept	228	2.40	0.33	175	2.38	0.25	53	2.49	0.51
Intrinsic value/Dispositional Interest	227	2.80	0.64	175	2.79	0.60	52	2.83	0.75
Attainment value	226	2.37	0.43	175	2.35	0.36	51	2.46	0.63
Utility value	227	3.56	0.54	175	3.54	0.54	52	3.62	0.55
Cost	227	2.45	0.62	175	2.40	0.55	52	2.62	0.80
Big five									
Conscientiousness	221	4.86	0.58	175	4.87	0.55	46	4.80	0.67
Extraversion	221	4.75	0.63	175	4.76	0.65	46	4.71	0.56
Agreeableness	221	4.85	0.62	175	4.86	0.62	46	4.80	0.63
Openness	222	4.86	1.15	175	4.85	1.15	47	4.91	1.19
Neuroticism	223	4.60	0.75	175	4.60	0.75	48	4.60	0.78
Present bias preferences									
Risk	222	0.68	0.20	175	0.68	0.20	47	0.69	0.20
Discount factor	217	0.98	0.68	175	0.94	0.55	42	1.14	1.04
Present bias	216	1.06	0.28	175	1.05	0.18	41	1.11	0.53
Achievement goals									
Mastery approach	219	6.15	0.71	175	6.12	0.74	44	6.25	0.61
Mastery avoidance	208	5.63	0.98	175	5.62	0.98	33	5.70	1.03
Performance approach	208	5.00	1.46	175	5.04	1.44	33	4.77	1.59
Performance avoidance	210	5.01	1.59	175	4.99	1.61	35	5.13	1.50
Subjective subject goals									
How many midterms?	223	2.81	0.46	175	2.82	0.46	48	2.79	0.46
How good in midterms?	223	0.79	0.13	175	0.79	0.14	48	0.80	0.12
Practice after midterms?	223	1.24	0.45	175	1.22	0.44	48	1.31	0.47
Which grade in exam?	223	2.05	0.62	175	2.05	0.62	48	2.03	0.62

Note: The table shows the number of observations, the mean and the standard deviation per variable for three different set of samples: first the raw sample in which we include all individuals who wrote the exam. The number of observations changes because some students did not answer the survey or did not answer some specific question of the survey. Next, we look at the complete-cases sample. There, we only included individuals for which we have all variables answered. Thereby, the number of observations is fixed for this sample for all variables. Lastly, we include the sample of incomplete-cases to show if our sample might differ due to the drop if individuals. Students who did not participate on the midterms, practice or MAD at all have zero points in the respective variable.

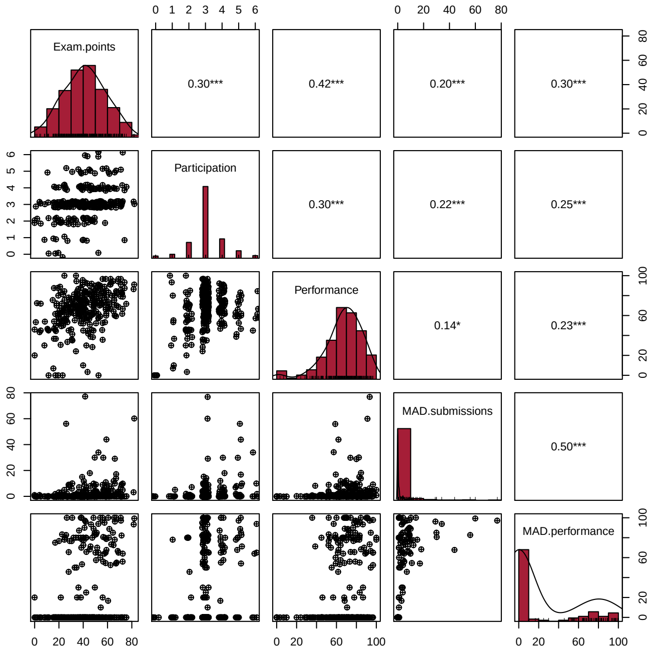
Overall, the exam appears to have been difficult as the average grade points are less than 45 out of 90. The maximum an individual student achieved was 82. However, it should be noted that for grading, the points in the midterm have to be added so that students got on average an additional 3 points so that the students in the complete sample pass on average. Our other key variables indicate that the average student participated in three midterms or practices and had nearly three times submitted a MAD solution.

As regards the sample composition, 56% of all students are female. Past performance in high school is rather good with an average high school GPA just above 2 and an average math grade of 2.6. 44% of our students seek one of the international degrees (B.Sc. International Business Administration or B.Sc. International Economics). The largest group of students with a minor in Business pursue a sports management degree.

The items from the expectancy-value theory source (source: Gaspard, Häfner, Parrisius, Trautwein & Nagengast, 2017, adapted to the university context and course), achievement goals (source: Elliot & Murayama, 2008, translated and adapted for the specific context), big five personality traits (source Schupp & Gerlitz, 2014, taken as is) and present bias preferences (source: Frederick, Loewenstein & O'Donoghue, 2002, translated) are without any extreme insights. Discussions and comparisons for the specific items can be found in the sources as well as in Marsh and Martin (2011), Wigfield and Eccles (2000), Hulleman, Schrager, Bodmann and Harackiewicz (2010), Loewenstein, O'Donoghue and Rabin (2003), Meier and Sprenger (2010), A. Becker, Deckers, Dohmen, Falk and Kosse (2012), Marsh et al. (2010). It is noteworthy that already at the beginning of the semester, the average of the aimed midterms was below three, and the number of practice runs is slightly above one. Lastly, on average, students set their grade goal for the exam equal to two in a grading system ranging from one to five.

Correlations, scatter plots, and distributions of the practice variables and the exam grade points are presented in Figure 5.3. All practice variables are positively correlated with the exam grade. There is also a strong correlation between participation and performance in the midterm tests and the matrix app. The univariate distribution of the respective variables is presented on the main diagonal plots in Figure 5.3. The exam points are close to a normal distribution. Participation in midterms and practice peaks at 3, which hints at most students taking the midterms only to earn extra points, but scarcely use them for additional practicing. In fact, about 2/3 of students with 3 midterms/practices did just the midterms. Performance in these midterms is heavily left-skewed, indicating

Figure 5.3 – Correlation plot



Note: The diagonal shows the distribution of the respective one-dimensional distribution. The lower half shows the two-dimensional scatterplot and the upper half the correlation.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

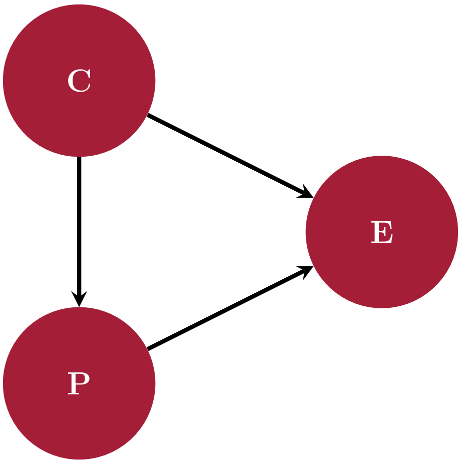
that students perform rather well in them. This is in stark contrast to the performance in the MAD, which is bimodal with a strong peak below 20%.

5.5 Model

In this chapter, we focus on the effect of practice on students' exam outcomes. Thus, we limit our analysis to four variables which capture students' practice behavior. More precisely, we have information on how often the students engaged in offers to practice, namely the number of midterm tests $participation_i$ and the number of submissions in the matrix app $MAD\text{-}submissions_i$. On the other hand, we also have measures of students' performance if the offers to practice were accepted, such as the mean points achieved in their trials $mean\ practice\ points_i$ and mean success rate of their handed-in solutions to the matrix app *MAD-percentages*$_i$.

To isolate a potential effect of practice P on the result in the final exam E, we control for a variety of other potential confounding factors C that include quantitative measures of students' personality, ability, goals, background and more. In essence, all factors covered by the control variables in our pool may affect both the exam outcome E and students' practice behavior P. As the control variables are predetermined, especially temporally, these factors cannot be affected by students' practice behavior. Figure 5.4 illustrates these relations schematically.

Figure 5.4 – Schematic structure of the variables



Note: The figure presents a schematic illustration of the relationship between the practice variables P, the control variables C and the result in the exam E.

As we only have a limited number of observations at hand, we have to carefully select which measures to include in the analysis, so that our results are not diluted by noise. Therefore, we employ the post-double selection method, introduced in Belloni, Chernozhukov and Hansen (2013), to limit the number of control variables and identify the treatment effect.

As a first step, we consider a model for the attained exam *points*$_i$ which does not control for any other potentially confounding factors:

$$\begin{aligned} \mathit{points}_i = \alpha &+ \rho_1\ \mathit{participation}_i + \rho_2\ \mathit{mean\ practice\ points}_i \\ &+ \rho_3\ \mathit{MAD\text{-}submissions}_i + \rho_4\ \mathit{MAD\text{-}percentages}_i + \varepsilon_i \end{aligned} \tag{5.1}$$

where ε_i is the unobserved individual specific error of the model. Clearly, in this model the practice variables are potentially endogenously determined. Considering the model in Equation (5.1) is, nonetheless, interesting in analyzing the effect of students' who did not use the practice opportunities and self-testing offers made available during the semester.

Individual factors that may bias the coefficients in the model in Equation (5.1) may include the motivation, reflected in personal goals or the ability, personality traits, income, etc. While there is no exogenous shock in the setup of this chapter that may help us to come closer to a causal interpretation of the coefficients, we have a rich set of control variables at our disposal that we put to use in the model specified in Equation (5.2). This enables us to exclude several of potentially confounding factors that may bias measurement of the direct effect of practice. Additionally, comparing estimation results of Equations (5.1) and (5.2) give insights by how much the practice effect might be biased not including the additional control variables.

In order to simplify notation, denote by $\mathbf{p}$ the (4×1) vector which contains the four regressors measuring students' practice behavior used in Equation (5.1). Then the fully specified model reads as follows:

$$\begin{aligned} \mathit{points}_i = \mu &+ \boldsymbol{\rho}'\mathbf{p}_i \\ &+ \boldsymbol{\beta}_1'\boldsymbol{char}_i + \boldsymbol{\beta}_2'\boldsymbol{EVT}_i + \boldsymbol{\beta}_4'\boldsymbol{agoals}_i \\ &+ \boldsymbol{\beta}_4'\boldsymbol{bigfive}_i + \boldsymbol{\beta}_5'\boldsymbol{pbp}_i + \boldsymbol{\beta}_6'\boldsymbol{sgoals}_i + \eta_i, \end{aligned} \tag{5.2}$$

where the index i stands for the individuals, and η_i is the idiosyncratic error term. $\boldsymbol{\rho}$, $\boldsymbol{\beta}_1$ through $\boldsymbol{\beta}_6$ are vectors of parameters with length determined by the number of factors included in each category. In $\boldsymbol{char}_i$ attributes of students such as their high school GPA, whether they took an advanced math class in high school, and their last math grade are subsumed. With these variables the background abilities of the students are captured. Measures from expectancy value theory measured in the survey and listed in Table 5.2 are represented with the vector $\boldsymbol{EVT}_i$, the personality traits captured by the survey, grouped into five factors, are included in the vector $\boldsymbol{bigfive}_i$ and the measures that identify present bias preferences are contained in $\boldsymbol{pbp}_i$. Achievement goals are enclosed in $\boldsymbol{agoals}_i$ while $\boldsymbol{sgoals}_i$ represents the subjective goals, i.e., the answers to the survey questions about students' self-set goals before the semester for the practices and the exam. For all variables descriptive statistics are presented in Table 5.2.

We are first and foremost interested in the coefficient of the practice variables. While, in principle, the practice variables themselves could be correlated with confounding variables, we do include measures for the most important aspects found by the literature on student achievement prediction to control for other confounders. With the help of our additional control variables, we can rule out confounding variables that relate to ability, self-concept, intrinsic value, attainment value, utility value, costs, personality traits, present bias preferences, achievement goals, and subjective goals for the class. Further, to account for different budget constraints, we include a dummy if students need to work to finance their studies. Clearly, we cannot completely rule out that we have omitted a variable of importance. However, given the rich set of controls and the prevailing literature at this point, we are confident that this chapter's setup is able to approximate a causal treatment effect of practice and self-control within the group of students analyzed.

We are also aware that this chapter is limited to students who participated in some way or another in the math lecture for economists in Tübingen. This means students selected themselves into an economics or business administration bachelor's degree program, for which the completion of the course is mandatory. While it happens that some students are surprised about the mathematical workload in their first semester, self-selection with regard to students' motivation plays only a subordinate role and is controlled for in our model set up.

In our analysis, we estimate Equation (5.2) model with a regression using basic OLS with heteroskedasticity robust standard errors. However, having only 175 students for which all variables are available, we face the situation that we have too few observations for the rich set of control variables.

This is why we turn to machine learning techniques for variable selection. With Lasso,[5] Random Forest and xgBoost, we employ three different techniques that are able to determine and select important features for the prediction of an outcome variable and are also able to rank their importance.

The shrinkage technique, first presented by Tibshirani (1996) and known as the Lasso, minimizes the sum of squared residuals subject to a constraint that penalizes the sum of coefficients. The Lasso estimate leads to a sparse representation of the model in Equation (5.2) since – due to the constraint on the sum of coefficients – some of the model parameters are forced to zero. The sparsity of the model is determined by the hyperparameter that controls the strength of the penalty on the coefficient sum. In order to select the value of the hyperparameter, we use a fine grid search across the hyperparameter space with a threefold repeated cross-validation with 999 repetitions. Our selection criterion is to minimize the root mean squared error ($RMSE$). In order to avoid over-fitting, we do, however, not directly select the value of the hyper-parameter that minimizes the $RMSE$ within all validations sets, but we add one standard deviation to it, which is common practice for the Lasso as suggested in Friedman, Hastie and Tibshirani (2001) and originally proposed by Breiman, Friedman, Stone and Olshen (1984). For this purpose, we use the unified interface provided through the `caret`-package in R (Kuhn, 2020) for machine learning techniques, as well as the methods developed by Friedman et al. (2001) provided in the R-package `glmnet`.

In addition, to the Lasso, we use the Random Forest technique to determine the most im-

[5] When we refer to the Lasso, it needs to be mentioned that we also analyzed elastic net specifications in each context. However, cross-validation results in all analyzed situations selected the pure Lasso specification.

portant variables. In essence, a Random Forest is the mean over the prediction of several regression trees. By aggregating the predictions of several trees, it reduces the variance of the predictions and avoids overfitting. This is also known as bootstrap aggregation (bagging) or bagged trees. Random Forest has one additional characteristic that distinguishes it from tree bagging algorithms: the selection of features is randomized at each potential split of the regression tree. Random Forest also allows inferring the variable importance by permuting so-called out-of-bag prediction errors. Random Forest goes back to Ho (1995), as well as Kleinberg (1996), and have been extended and trademarked by Breiman (2001). In our analysis, we rely on the R-packages `randomForest` developed by Liaw and Wiener (2002) and `Boruta` written by Kursa and Rudnicki (2010). Different from Lasso, which imposes a linear functional structure between outcome and regressors, Random Forest is similar to multivariate kernel methods. As such, variables identified by a Random Forest as important may be related in a nonlinear functional relation to the outcome variable. For feature selection, we use the algorithm provided in the `Boruta`-package, where the variable importance based on the out-of-bag prediction error of features is repeatedly competing against the out-of-bag error of a reshuffled version of the feature, the so-called shadow features, in order to detect whether a certain feature outperforms its shadow feature significantly. If this comparison is done sufficiently often, given a certain significance level, the Boruta-algorithm decides based on a test statistic whether a certain feature is helpful for the description of the outcome variable or not. In our case, we set the maximal number of comparisons to 9,999 and chose a significance level of 5%.

Nonetheless, since features are selected randomly when building the trees in the Random Forest, the importance of a feature is split across several highly correlated variables. In essence, the strongly correlated features will end up with about the same variable importance. However, their importance is reduced compared to the situation where only one feature is included in the analysis. Also, if many highly correlated variables are present that contain valuable information on the outcome variable, Random Forest will select the information contained in the correlated features more often while it may neglect other valuable information.

Therefore, as a third method, we use Extreme Gradient Boosting (xgBoost, T. Chen et al., 2020), which is another method that relies on decision trees. Instead of simply taking the mean over several decision trees, xgBoost uses gradient boosting to combine several regression trees. xgBoost also contains elements of regularization and shrinkage. Usually, the depth of the regression trees used in the sequential boosting procedure is not very deep. The advantage of xgBoost in comparison to Random Forest is that among highly correlated features, xgBoost selects only one of the features since, with the sequential boosting, features are incrementally added to the model. If one feature is selected early, other highly correlated features can improve the prediction only if they bring new information to the table. In our application, the tree depth is selected via a grid search over the hyper-parameter space in a threefold repeated cross-validation with 9,999 repetitions. Also, the learning rate, the regularization, and shrinkage parameters as well as other parameters, are determined in this fashion.

After the variable selection via the three machine learning methods, we follow Belloni et al. (2013) and run post-selection OLS regressions. It is important to note that we do not only select variables for our respective outcome, but also for explanatory variables (here: the four practice variables). Therefore, the double refers to the two-step or, in our case, five-step selection process. One not only gets the most predictive variables for the model itself, but also the variables that help to reduce a possible bias of our explanatory variables of interest.

So far, the models discussed in Equations (5.1) and (5.2) are designed to answer the question whether there is an effect of practice and self-testing on exam outcomes, and, if yes, in which direction and how strong it is. With the relatively small number of observations at hand, forming groups based on several variables and dissecting the coefficient $\boldsymbol{\rho}$ into group specific effects is not of great avail. Therefore, we also turn in our last model to the method of quantile regression (Koenker & Bassett, 1978), to analyze how the effect of practice varies across performance groups. The post-selection quantile regressions allow us to give cursory answers to questions like: do students who end up in the top-performing group share certain characteristics? What about students in the lower performing groups? Are there diminishing returns on practice? Who benefits the most from practice?

Following Koenker and Bassett (1978), the conditional quantile model for exam grade points is specified as

$$Q_{Points_i}(\tau|\mathbf{p}_i, \boldsymbol{x}_i) = \alpha_\tau + \boldsymbol{\rho}'_\tau \mathbf{p}_i + \boldsymbol{\beta}'_\tau \boldsymbol{X}_i \tag{5.3}$$

for quantiles $\tau \in (0, 1)$ in steps of 0.05 In contrast to Equation (5.2), quantile regression specifies a linear model for every quantile of the dependent variable $points_i$. Hence, the parameter estimates $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ may vary across τ's. To keep notation short, $\boldsymbol{X}_i$ contains all control variables.

5.6 Results

As a first step of our analysis, it is useful to consider the results of an uncontrolled regression of the exam points on the various practice parameters only (corresponding to Equation (5.1). These results are presented in Table 5.3.

In this setting, as only the four practice variables enter the equation, we can use the full set of 280 students who have taken the exam. Moreover, in a second step, we can then compare the coefficients of the variables for students that have not completed the survey and for whom we do not have a full set of control variables available. This makes it possible to infer the direction of a possible selection bias.

Comparing the specifications (1) and (2) in Table 5.3 which both use the full sample of 280 observations, the coefficient estimates for the participation variables (number of taken midterms/practices and submissions to MAD) decrease when the performance of the practice sessions and submissions enter the model. To give some context to the size of the coefficients, the coefficients in (1) imply that a student who – like the majority of students – participated in the three midterm test is, on average, expected to obtain an

Table 5.3 – Sequential inclusion of practice variables

	Dependent variable: Points in end exam					
	(1)	(2)	(3)	(4)	(5)	(6)
Practice participation	4.619***	2.607***	5.698***	4.633***	2.106	3.495**
	(0.945)	(0.983)	(1.367)	(1.229)	(3.570)	(1.434)
Mean points per practice		0.297***		0.365***	0.500***	−0.027
		(0.056)		(0.068)	(0.142)	(0.089)
MAD-Submissions	0.284*	0.066	0.391*	0.144	−0.059	1.176***
	(0.163)	(0.154)	(0.235)	(0.235)	(0.249)	(0.396)
MAD-Percentages		0.079***		0.064*	0.092	−0.029
		(0.028)		(0.037)	(0.061)	(0.094)
Constant	26.116***	11.096**	24.103***	0.546	−2.931	25.094***
	(3.086)	(4.377)	(4.586)	(5.723)	(13.160)	(5.298)
Adjusted R^2	0.103	0.235	0.113	0.234	0.312	0.151
Observations	280	280	175	175	53	52

Note: Column (5) includes individuals who only partically answered the survey. Column (6) includes individuals who did not take part in the survey at all. Heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

additional 13 points in the final exam compared to a student who did not take any of the opportunities. In this prediction, the extra points the student could have earned with him/her successful submission to the midterms are excluded. In addition, the exam prediction for a student who submitted 10 matrices on different days is predicted to have roughly 3 extra points. Note that, on average, students who participated in the MAD-app submitted around 8 solutions. One student in our sample submitted up to 77 solutions to MAD.

When the performance of the submissions or midterm/practice is considered as well (specification 2 in Table 5.3), we see that the coefficients of the participation variables decrease. The number of submissions to the MAD-app becomes insignificant. Simultaneously, the performance variables take up some of their shares and are able to capture additional variation in the outcome variable. This first result would suggest that both practice activity as well as the performance during practice may serve as predictors of the final exam outcome. In the following, we will analyze how robust this result is when the sample is restricted.

As a first robustness check, we only consider the 175 students for which we have a full set of sensible survey entries in the framework of the reduced model in Equation (5.1). The results are presented in columns (3) and (4) in Table 5.3. Again, we first include the practice participation measures and then add the performance measures. We find that the coefficients of the participation measures also decrease when performance is added. More interesting, however, is the increase in the coefficients of the measures related to the midterms as well as the increase in the coefficient of MAD submissions. With the exception of the MAD-app performance, this result indicates that in the reduced sample of 175 full observations, the effect of practice may be slightly overestimated. The reduction in the coefficients for the performance in the MAD-app indicates the contrary, i.e., that the effect is underestimated in the reduced sample. However, comparing models (1) and (3), we would not be able to reject the hypothesis that the coefficients are equal.

The comparison between the full sample and the reduced sample of full cases can be detailed further with the results in columns (5) and (6) of Table 5.3. The observations missing in the reduced sample of the 175 full cases have either not taken the survey at the beginning of the lecture or gave incomplete information on several survey questions. Both groups comprise around the same number of students. In the fifth column of Table 5.3, we perform a regression of the exam points on all practice variables for those 53 students who only partially answered the survey and given incomplete information. On a side note, the pattern in the missing answers (often answers to the last questions are missing), these students seem to be easily bored or have exerted some degree of laziness or inadvertence when filling out the survey. For those students, the performance during practice is especially important in order to predict the final exam outcome. So, if those students were able to work in the midterm exams meticulously, they were more likely to achieve a higher score in the exam.

For the other group of 52 students who did not take the survey at all, the regression results suggest that participation trumps performance. Recall this group comprises 23 students who were registered for the exam in former semesters, and who had to retake the entire lecture together with the exam. The overall distribution of final exam results in this group is also stochastically dominated by the outcome distribution in the group of students who gave only partial answers in the survey. Participation rates among these students are also lower. Only around 60% of these students did participate in at least three midterm/practice tests while 0.83% of the students that gave partial answers in the survey and 0.89% of the students that gave full answers. Also, the number of submissions to the MAD-app is reduced in this group. Nonetheless, those who participated benefited from it.

In conclusion, our first impression is that we slightly over-estimate the effect of practice and self-testing in the sample of 175 students who have provided full information, compared to the full sample. These students may have, in general, a higher willingness to engage and, thus, are more receptive to the benefits of the practice and self-testing opportunities provided in the course of the lecture. However, this personality trait is controlled for by the variables in the survey that specifically asks for information that captures students' motivation and goals towards the course. Therefore, we deem the bias introduced by the reduction of the sample size to 175 complete cases immaterial.

Therefore, we now turn to the model framework setup in Equation (5.2) for which we only use these 175 complete cases. Table 5.4 presents a summary of the results for regressions with different sets of control variables. In the columns (1) to (6), each vector of control variables from Equation (5.2) is included separately. Column (7) then presents the result for a regression where all control variables are jointly included.

The major finding across all specifications is that the estimates of participation, submission to the MAD, and performance in the MAD are rather stable. For example, the estimate of participation varies between 4.273 and 5.044. Considering the standard deviation, we would again not be able to reject a hypothesis that the estimates are equal across specifications. In contrast, the performance during practice is only weakly significant in the models that control for individual characteristics (columns 1 and 7). When we introduce the psychology measures, the effect is estimated higher with a maximum of 0.367 in column (2). Hence, individual characteristics seem to be more relevant for the final exam performance than the performance in the practice opportunities during the

Table 5.4 – Sequential inclusion of practice variable (complete-cases subsample)

	Dependent variable: Points in end exam						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Practice participation	4.273^{***} (1.164)	4.609^{***} (1.268)	4.788^{***} (1.362)	4.621^{***} (1.224)	4.489^{***} (1.245)	4.868^{***} (1.199)	5.044^{***} (1.357)
Mean points per practice	0.135^{*} (0.072)	0.367^{***} (0.068)	0.357^{***} (0.072)	0.354^{***} (0.069)	0.343^{***} (0.066)	0.347^{***} (0.073)	0.139^{*} (0.077)
Submissions MAD	0.174 (0.157)	0.147 (0.225)	0.106 (0.231)	0.116 (0.248)	0.245 (0.229)	0.064 (0.189)	0.123 (0.193)
Percentages MAD	0.055^{*} (0.031)	0.063^{*} (0.037)	0.076^{**} (0.037)	0.069^{*} (0.037)	0.062^{*} (0.036)	0.078^{**} (0.035)	0.065^{*} (0.033)
Individual characteristics	Yes	No	No	No	No	No	Yes
Expectancy-value theory	No	Yes	No	No	No	No	Yes
Big five personality traits	No	No	Yes	No	No	No	Yes
Present risk preferences	No	No	No	Yes	No	No	Yes
Achievement goals	No	No	No	No	Yes	No	Yes
Subjective subject goals	No	No	No	No	No	Yes	Yes
Number of coefficients	15	10	10	8	9	9	36
Adjusted R^2	0.455	0.225	0.240	0.238	0.261	0.284	0.448
Observations	175	175	175	175	175	175	175

Note: The set of control variables is described in Table 5.2. Heteroskedastic robust standard errors in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

semester. The performance effect is about halved when the ability measures in column (1) are included. This could be seen positively as students do have the chance to improve during the semester, starting out with weak performances in the midterms and still achieving a good grade.

The result also seems to mirror the importance of practice found in the musical practice literature, for example, Sloboda, Davidson, Howe and Moore (1996), who found that high achievers are more consistent in their practice patterns. Simultaneously, Sloboda et al. (1996) did not find that high-achievers "can gain a given level of examination success on less practice than low achievers". Formal musical practice is much like constant self-testing, since direct feedback and errors are immediately apparent, especially when learning an instrument. The study of Sloboda et al. (1996) culminated in the finding that, with hefty individual differences, it takes an average of 3,300 hours to achieve the highest grade level in their study. Sloboda et al. (1996) found that mostly participation in regular practice leads to successful learning. As direct feedback and the possibilities to make errors are also important aspects of the testing setup in our analysis, this is what we also find: self-testing during the semester has enhanced retention of learned materials in the final exam, regardless of the performance during the tests. This is also in line with the research conducted in the line of Roediger III and Karpicke (2006). Additionally, we find that performance during practice matters as well for the final exam outcome or at least is a good predictor for the final outcome.

With regard to the practice variables, nonetheless, a note of caution is due: the variables that measure mere practice participation in our case may also capture another effect that echos an observation described in Clark and Bjork (2012): "Students can be learning even when a current test shows little or no progress". Students regulate what to learn in the future, guided by their test outcomes. The results presented in Table 5.4 for the participation variables thus may also capture effects like the "hyper-correction effect" documented by Butler, Karpicke and Roediger III (2008) in which errors in pre-tests made with high confidence are prominently corrected in final tests. But also correct answers given with low-confidence are enforced. These correction effects are, thus, somewhat independent of the performance in the respective tests.

The adjusted R^2 increases from 0.445 in (1) to only 0.448 in (7). This suggests, together with the rather low number of significant coefficient estimates, that the model in (7) includes too many variables and a more sparsely designed model may provide just as much explanatory content. Looking at the adjusted R^2, it is also noteworthy that when individual characteristics are not accounted for, the adjusted R^2 drops to values around 0.23, highlighting the importance of these variables to explain exam outcome.

5.6.1 Variable selection regression results

As discussed above in Section 5.5, to avoid overfitting, we employ three methods to select relevant predictor variables in addition to the practice variables: Lasso, Random Forest, and extreme gradient boosting (xgBoost). We first estimate a model with the exam and the four practice variables to determine the major covariates which explain these variables. Each of the three methods selects a set of variables important for the prediction. In the last step, we re-estimate a sparse alternative to the model in Equation (5.2) using each variable set separately in addition to the practice variables. The results of the regressions are presented in the Table 5.5.

Table 5.5 – Post double selection OLS regressions

	Dependent variable: Points in end exam		
	(1) Lasso	(2) Random Forest	(3) xgBoost
Practice participation	3.692***	4.200***	4.130***
	(1.048)	(1.267)	(1.343)
Mean points per practice	0.161**	0.163**	0.169**
	(0.072)	(0.070)	(0.074)
Submissions MAD	0.143	0.188	0.143
	(0.194)	(0.179)	(0.457)
Percentages MAD	0.059*	0.071**	0.061
	(0.031)	(0.030)	(0.040)
Number of Coefficients	12	18	16
Adjusted R^2	0.464	0.452	0.420
Observations	175	175	175

Note: Column (1) shows the post-double selection results using the Lasso algortihm, column (2) using the Random Forest and column (3) using the xgBoost. For all three columns, we use a double selection process, in which we select variables that are important for the (i) the outcome and (ii) the practice variables. The selection is done by the named algorithm. Heteroskedastic robust standard errors in parentheses. The full set of estimated coefficients are include in Table D1. $^+$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Measured by the adjusted R^2, the sparse models have approximately the same explanatory content as the full model above. Only the model selected by xgBoost achieves a slightly lower R^2. This is mainly due to the omission of the students' math grades in high school, which is highly correlated with the overall high school GPA. In a linear regression model, the two variables add extra information to the model. However, xgBoost is agnostic to the functional structure and is also sensitive to non-linear interactions of the variables. Additionally, as explained above in Section 5.5, it selects only one of a few highly correlated variables. This favors variables of relevance that may have otherwise been neglected in the prediction. This can directly be seen in Table 5.5. Among the linear models considered in Table 5.5, Lasso selects the most spares and also best fitting model.

The results for the sparse models in Table 5.5 also shows that the coefficients for the practice variables are rather robust. We cannot reject the null hypothesis that the respective magnitudes of the effects measured in the sparse models are the same as measured for the full model in column (7) of Table 5.4.

By means of machine learning techniques, we have, thus, shown that the full model Equation (5.2) can approximately be represented with the sparse variable sets presented in Table 5.5, while simultaneously maintaining the effect size of the practice variables.

5.6.2 Quantile regression

In order to shed more light on the question who benefits from the practice, we present subsequently the results of the quantile regression model specified in Equation (5.3). The covariates in $\boldsymbol{X}_i$ are either specified as the union or the intersection of all relevant covariates identified in Section 5.6.1. The results are graphically displayed in Figures D1 to D4, restricted to the practice variables of interest. Confidence intervals of the quantile regression coefficients are based on standard errors calculated using a wild bootstrap procedure with 5,000 replications.

Figure D1 presents the coefficient estimates on the number of midterm tests. In both specifications (a) and (b), the parameters are usually statistically significant with some exceptions for quantiles around 0.8 in (a). Regarding Figure D1a, we can see that students who are expected to achieve a low number of points in the exam, i.e., the lower quantiles, benefit more from the practice than students who are expected to attain a high number of points anyway. For the lowest quantile, additional practice adds 6 points to the final exam while the effect goes down to roughly 2 points for students in the upper quantile. When using all control variables (see Figure D1b), this pattern is less pronounced. The lower quantile estimates still are as high as before, but the upper quantile estimates vary now around 4 as well. Hence, practicing is helpful in general, and it seems to be the case that weak students benefit even more than otherwise good students.

The effect of the precise performance in the midterm tests is less clearcut. Considering

Figure D2a, we see that we have a statistically significant relationship in the very low quantiles (below 20%) which suggests that good performance in the midterm tests is an indicator for better performance for very weak students who are expected to achieve a low number of points. However, the effect vanishes (in both a and b) for the 20 to 40% quantiles. For the median to 80% quantile students, good performance during the midterm also predicts a higher number of points in the final exam. For the very good students (upper quantiles), there is no significant relationship. Hence, irrespective of their midterm test outcome, these students are expected to achieve a high number of grade points in the final exam which is plausible.

How often a student made use of the MAD app does not seem to have an effect in the specification presented in Figure D3a, as suggested by the regression results before. This holds across all quantiles. In contrast, Figure D3b might imply a slightly negative relationship between participation in the MAD and exam performance for the very weak, i.e., the lowest 5% quantile.

Nevertheless, good performance in the MAD appears to be a good predictor for successful exam outcome, in particular for students who are expected to achieve only a low number of points in the exam. This holds for both specifications presented in Figures D4a and D4b.

5.7 Conclusion

Our analysis evaluated several e-learning exercises during one semester accompanying the math-lecture designed for students of business and economics studies in their first semester. We found that positive learning gains are associated with students' participation in these exercises. The participation in the midterm/practice tests accounts for a sizable increase in the points attained in the final exam. The number of submissions to the matrix app (MAD) was, however, insignificant. In addition to the participation effect, we find that students who had a higher performance in the e-learning opportunities (MAD as well as midterm/practice tests) were more likely to obtain higher scores in the final exam.

In our analysis we employ a rich set of control variables in order to approximate the causal effect of practice participation and performance on the exam points. We employ several robustness checks with regard to the selection of students as well as overfitting. To counter the concern of overfitting, we employ variable selection techniques from the field of machine learning, namely Lasso, Random Forest, and xgBoost. In all cases, the different sets of selected variables did not change the measured effect of practice performance and participation significantly.

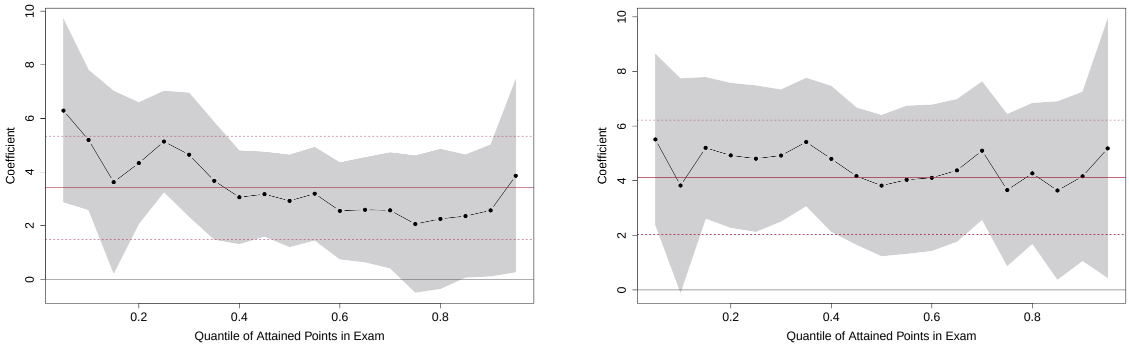
Finally, we used quantile regression of final exam points on the most important variables identified by the machine learning techniques to explore the question of how much the various performance groups in the final exam have benefited from practice and self-testing. We show that especially the students with lower points benefited the most from additional practice.

Therefore, our results suggest that giving students the possibility to self-test and practice the material in online settings with knowledge of correct response helps students to improve (math) exam grades.

Appendix

D.1 Figures

Figure D1 – Quantile regression coefficients: Number of taken midterm tests

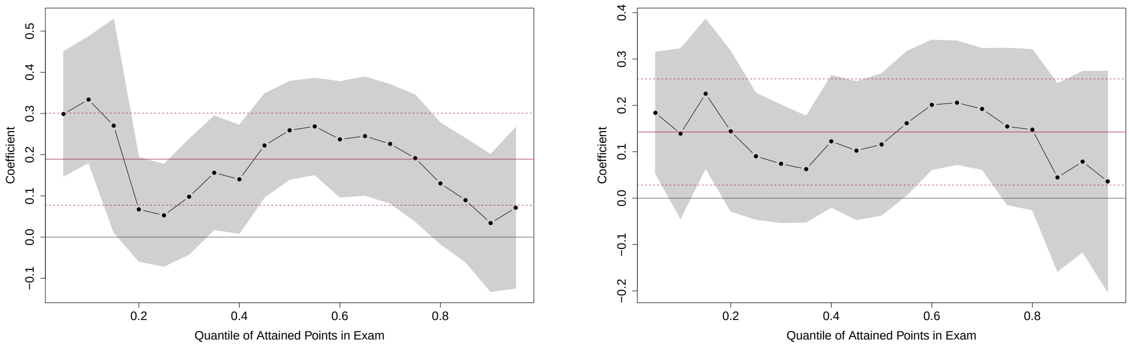


(a) Intersection

(b) Union

Note: The panel shows the estimates for the coefficients of the number of midterm tests taken across quantiles. The quantile specific estimates are obtained from a quantile regression on the points obtained in the exam. Figure D1a presents the coefficients when the intersection of all variables selected by machine learning techniques are considered in Section 5.6.1. Figure D1b presents the coefficient when the union of all selected variables comprises the control set. The solid red horizontal line shows the value obtained in an equivalent OLS regression, while the dotted red lines indicate the 90%-confidence bounds of the OLS estimate. The shaded areas identify the 90% confidence bounds of the quantile regression estimates. Standard errors of the quantile regression coefficients have been calculated based on a wild bootstrap procedure with 5000 replications.

Figure D2 – Quantile regression coefficients: Obtained points in midterm test

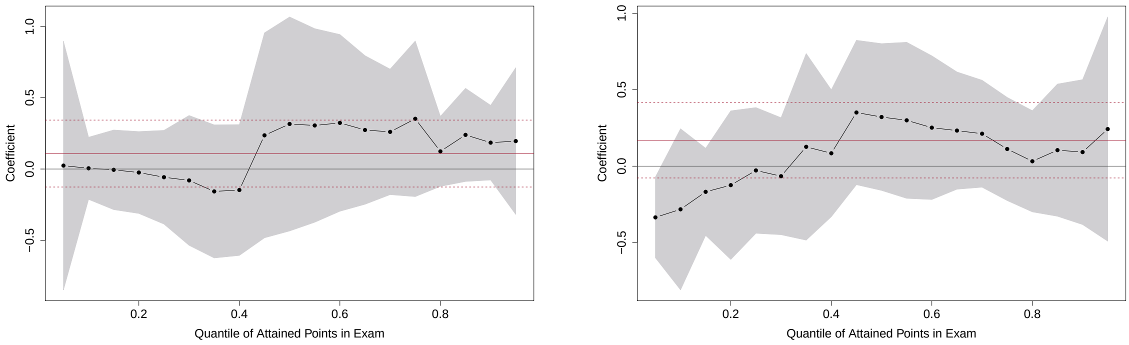


(a) Intersection

(b) Union

Note: The panel shows the estimates for the coefficients of the mean achieved points in the midterm tests across quantiles. The quantile specific estimates are obtained from a quantile regression on the points obtained in the exam. Figure D2a presents the coefficients when the intersection of all variables selected by machine learning techniques are considered in Section 5.6.1. Figure D2b presents the coefficient when the union of all selected variables comprises the control set. The solid red horizontal line shows the value obtained in an equivalent OLS regression, while the dotted red lines indicate the 90% confidence bounds of the OLS estimate. The shaded areas identify the 90% confidence bounds of the quantile regression estimates. Standard errors of the quantile regression coefficients have been calculated based on a wild bootstrap procedure with 5000 replications.

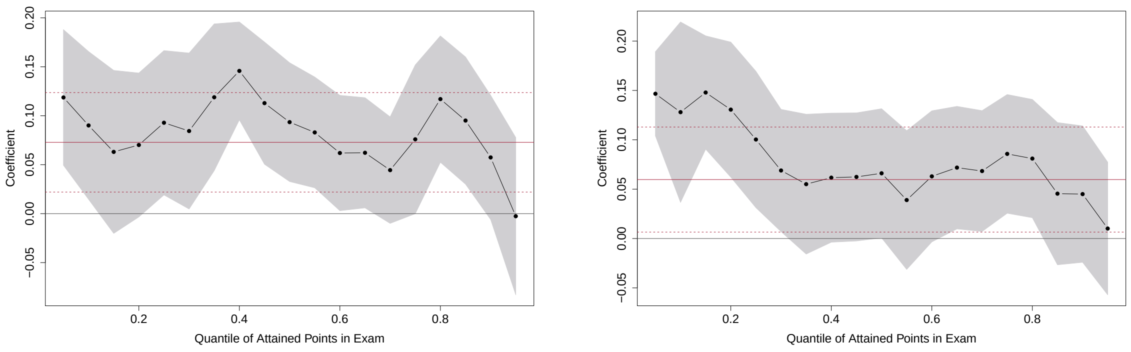
Figure D3 – Quantile regression coefficients: Number of submissions to MAD



(a) Intersection **(b)** Union

Note: The panel shows the estimates for the coefficients of the number of submissions to the matrix app across quantiles. The quantile specific estimates are obtained from a quantile regression on the points obtained in the exam. Figure D3a presents the coefficients when the intersection of all variables selected by machine learning techniques are considered in Section 5.6.1. Figure D3b presents the coefficient when the union of all selected variables comprises the control set. The solid red horizontal line shows the value obtained in an equivalent OLS regression, while the dotted red lines indicate the 90%-confidence bounds of the OLS estimate. The shaded areas identify the 90% confidence bounds of the quantile regression estimates. Standard errors of the quantile regression coefficients have been calculated based on a wild bootstrap procedure with 5000 replications.

Figure D4 – Quantile regression coefficients: Achieved percentages MAD



(a) Intersection

(b) Union

Note: The panel shows the estimates for the coefficients of the mean percentage of correct answers for submitted solutions to the matrix app across quantiles. The quantile specific estimates are obtained from a quantile regression on the points obtained in the exam. Figure D4a presents the coefficients when the intersection of all variables selected by machine learning techniques are considered in Section 5.6.1. Figure D4b presents the coefficient when the union of all selected variables comprises the control set. The solid red horizontal line shows the value obtained in an equivalent OLS regression, while the dotted red lines indicate the 90%-confidence bounds of the OLS estimate. The shaded areas identify the 90% confidence bounds of the quantile regression estimates. Standard errors of the quantile regression coefficients have been calculated based on a wild bootstrap procedure with 5000 replications.

D.2 Table

Table D1 – Full regression results of Table 5.5

	Dependent variable: Points in end exam		
	(1) Lasso	(2) Random Forest	(3) xgBoost
Practice participation	3.692***	4.200***	4.130***
	(1.048)	(1.267)	(1.343)
Mean points per practice	0.161**	0.163**	0.169**
	(0.072)	(0.070)	(0.074)
Submissions MAD	0.143	0.188	0.143
	(0.194)	(0.179)	(0.457)
Percentages MAD	0.059*	0.071**	0.061
	(0.031)	(0.030)	(0.040)
HS GPA	−10.755***	−9.625***	−13.842***
	(2.271)	(2.299)	(2.158)
Last math grade	−3.155***	−3.091***	
	(1.106)	(1.090)	
Sport studies	9.162**		
	(4.413)		
Second field students	−0.693	−2.628	
	(3.744)	(3.735)	
Work to finance studies	−2.716		
	(2.773)		
Mastery Approach (AG)	2.426*		
	(1.290)		
Risk (PBP)	8.009	9.764*	11.851**
	(4.901)	(5.512)	(5.832)
Retaker		−11.690**	
		(4.871)	
Number of semesters		−0.928	
		(0.738)	
International studies		−3.648*	−3.195
		(2.172)	(2.434)
Performance Approach (AG)		1.594	0.251
		(1.233)	(0.823)
Mastery Avoidance (AG)		2.435**	2.131*
		(1.166)	(1.273)
Performance Avoidance (AG)		−1.503	
		(0.999)	
Extraversion (BF)		−1.183	
		(1.555)	
Agreeableness (BF)		−1.004	−1.533
		(1.761)	(2.002)
Present bias (PBP)		1.361	1.630
		(5.343)	(7.681)
Female			−2.259
			(2.551)
Openness (BF)			−0.517
			(1.127)
Neuroticism (BF)			−1.336
			(1.640)
Discount Factor (PBP)			−0.977
			(1.448)
Constant	28.496**	58.622***	41.485***
	(12.710)	(16.341)	(15.450)
Adjusted R^2	0.464	0.452	0.420
Observations	175	175	175

Note: The table show the complete estimation results of Table D1. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 6

Practice makes perfect? Evidence from a voluntary self-testing e-learning setting*

*This chapter is based on: Schwerter, J.(2020): Practice makes perfect? Evidence from a voluntary self-testing e-learning setting, unpublished manuscript, University of Tübingen.

6.1 Introduction

Recently, e-learning tools, video teaching, and similar teaching forms have gained significantly more relevance within higher education teaching. This relevance will increase even further due to the worldwide pandemic in 2020. The academic literature evaluating this new way of teaching has been growing accordingly (for example, Broadbent & Poon, 2015; B. W. Brown & Liedholm, 2002; M. G. Brown, 2016; Butler, 2010; D. Coates et al., 2004; Figlio et al., 2013; Fischer, Zhou et al., 2019; Kizilcec et al., 2017; O'Flaherty & Phillips, 2015; Paechter et al., 2010; Thai et al., 2017; D. Xu & Jaggars, 2014). Most of the studies focus on the comparison of face-to-face teaching, blended teaching, and video (or online) teaching. The present study focuses on a different aspect of online teaching, namely additional voluntary e-learning practices. I examine whether online practice helps students prepare for the exam while controlling for the main student achievement predictors such as motivation, ability, and self-concept.

To this end, I observed students over a whole semester in their participation in an e-learning environment, accompanying the course *Social Science Statistics II*. This course is for third-semester bachelor students from the University of Tübingen, Germany. I measured students' practice behavior counting the number of weekly e-learning sessions they participated in, the mean performance in these sessions, and the mean number of trials per session. Further, to account for the general engagement of the students with the topics, I survey students' weekly mandatory preparation for the face-to-face tutorial. My results show that participation in the e-learning sessions, independent of performance, lead to better exam grades. Further, being better prepared for the face-to-face tutorial also leads to a higher score in the exam. The estimation results are robust to controlling for ability, motivation, achievement goals, personality traits, present bias preferences, and subjective goals.

The chapter is structured in the following way: first, I refer to the literature the chapter is connected to in Section 6.2. Then, I give an overview of the course structure and the e-learning environment in Section 6.3, and the data in Section 6.4. Section 6.5 shows the estimation model while Section 6.6 presents the results. The conclusion follows in Section 6.7.

6.2 Literature

The chapter contributes to the rich literature on the prediction of exam grades. McKenzie and Schweitzer (2001) showed that for first-year Australian University students, their previous academic performance, integration into university, self-efficacy, and employment responsibility are essential predictors for exam grades. Paechter et al. (2010) add that achievement goals and students' motivation is of high importance for students' success. This is confirmed by Komarraju and Nadler (2013), using 407 psychology students in the USA, and adding effort regulation and help-seeking behavior. Bailey and Phillips (2016) confirm for first-year students the high positive impact of intrinsic motivation. Rimfeld et al. (2016), among others, emphasize the importance of personality measures such as the big five, and that grit, i.e., perseverance and passion for long-term goals, adds only marginal explanation for exam achievements.[1] Within my study, I gather those high predictive variables to control for important confounders for students' achievement.

Condron, Becker and Bzhetaj (2018) show that about two-thirds of the sociology, anthropology, social work, and criminal justice students at a public university in the midwest of the USA have high statistics anxiety. They further demonstrate that confidence is one of the primary drivers of statistics anxiety. Thus, the group of sociology students in my sample should be of particular interest to examine the help of the e-learning practice. Thus, I add to the specific statistics literature on who succeeds in exams and how students' exam outcomes can be improved. I do not only survey psychology measures and see if these explain exam grades. I also add to the e-learning literature by including the voluntary e-learning exercises with which students can practice during the

[1] See Honicke and Broadbent (2016) for a more comprehensive review on this topic.

semester. In this way, I also contribute to the self-testing literature. Rodriguez et al. (2016), Rodriguez, Kataoka et al. (2018), and Rodriguez, Kataoka et al. (2018) underline the importance of spacing and self-testing in learning and their positive influence on students learning achievement.

J. Park et al. (2018) confirm the results on spacing. Using clickstream data, procrastinators perform significantly worse in exams. This group benefits from higher levels of regularity, while it is the other way around for non-procrastinators. Baker et al. (2019) aim to improve student's time management by letting students scheduling their online lectures themselves. Students who were allowed to schedule it on their own were better in the first quiz, especially those with low reported time management skills. This effect, however, did not persist over time and did not lead to better exam grades. The intervention did not show any effect on students' behavior in terms of cramming, procrastination, or the time at which students studied. Fischer, Zhou et al. (2019) present results for a three weeks online preparation course in chemistry. They show participation improved exam grades by one-third of a letter grade. Especially at-risk students benefitted from online preparation. Thus, my results add to the question of spacing, participate, and performance of self-testing (or practicing).

6.3 Course information and e-learning environment

6.3.1 Course information

The general course structure is summarized in Table 6.1. The lecture spans over fifteen weeks with thirteen lectures. The lecture is accompanied by weekly tutorial sessions with mandatory attendance during the week in which tutors present solutions to the problem sheets to the students. The students had to choose the date of one out of four of these tutorials at the beginning of the semester.[2] There were two additional sessions, one in the middle of the semester and one at the end, in which previous exam questions were solved. 20-30 students attended each of the tutorials. If they missed more than two sessions, they were not allowed to write the exam at the end of the semester. Thereby, the requirement for passing the tutorial is to be present; it is up to the students whether they prepared the respective problem sheets discussed during the tutorial, and whether they actively participate.

Table 6.1 – Semester structure

Week	Date.of.lecture	Topics	Tutorial
1	2019-10-14	(1) Probability	Pretest and survey
2	2019-10-21	(2) Discrete random variables	Exercise sheet 1
3	2019-10-28	(2) Continuous random variables	Exercise sheet 2
4	2019-11-04	(3) Specific discrete distributions	Exercise sheet 3
5	2019-11-11	(3) Specific continuous distributions	Exercise sheet 4
6	2019-11-18	(4) Twodimensional distributions	Exercise sheet 5
7	2019-11-25	(5) Theorems and sample mean	Exercise sheet 6
8	2019-12-02	(6) Point estimation	Exercise sheet 7
9	2019-12-09	(7) Interval estimation	Exercise sheet 8
10	2019-12-16	(8) Statistical testing and p-value	Exercise sheet 9
NA	2019-12-23	Christmas break	Christmas break
NA	2019-12-30	Christmas break	Christmas breask
NA	2020-01-06	National holiday (affected only the lecture)	Old exam questions
11	2020-01-13	Rep. (6), (7) and (8), and introduction of (9)	Exercise sheet 10
12	2020-01-20	(9) Regression analysis	Exercise sheet 11
13	2020-01-27	(9) Regression analysis, examples	Exercise sheet 12
14	2020-02-03	Question session	Old exam questions
15	2020-02-10	Exam	

Note: The tutorials were held on Wednesday and Thursday of the same week of the corresponding lecture. Week 11 includes a repetition of topics (6), (7) and (8) to show the connection of the topics and why they are important for topic (9). Every exercise sheet had an additional e-learning exercise. Only after week 12, there was an additional e-learning exercise session for Stata-Questions. Thus, there were thirteen e-learning exercises.

If students already visited the tutorial with less than two absences in another semester in the years before, attending the tutorial was voluntary. This was usually the case for 'retakers'. At the beginning of these face-to-face tutorials, students had to sign a list to prove attendance. When students signed the list in the tutorial, I additionally asked

[2] Selection was based on the concept *first come first serve.*

them to answer whether they (i) did nothing to prepare themselves for the tutorial, (ii) had a look at the exercises before the tutorial, (iii) tried to solve the exercises, or (iv) completely solved the exercises to the best of their knowledge, allowing for intermediate steps.

In addition to the tutorial, the solutions to the problem sheets were pre-recorded in videos from the previous semester. Half of the exercises each week had the exact same numbers as in the videos, the other half of the exercises have different numbers in the exercises, while the frame was the same. It is left to the students to decide when, or even whether, they watch the videos. Due to a technical problem, I could not observe who had watched the videos. Then, as the center of the study's design, students can practice the week's topic with the help of e-learning exercises. These exercises covered between one to three exercises of the weekly tutorial, consisting of the same frame or wording as those in the tutorial, but with new examples. The number of exercises depends on the respective length and difficulty. the course, including the topics and the date of the exam.

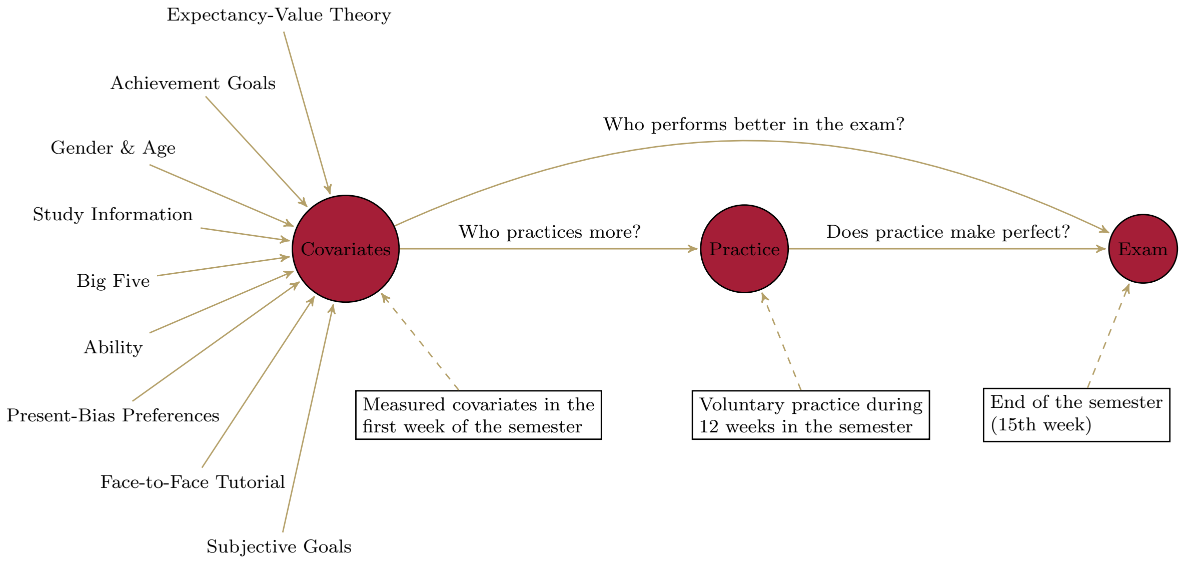
6.3.2 Design of e-learning exercises

This study aims to examine if additional practice helps to achieve better grades in *Social Sciences Statistics II* (called only Statistics 2 from now on), as pictured in Figure 6.1. The additional e-learning exercises were provided weekly and voluntarily with direct feedback on the online management system of the university.[3] There is no personalized feedback, but the students got the information whether their individual answers were correct or false, and the correct solution for the latter. Additionally, students saw how many points they achieved in total at the end of the exercises. There is no personalized feedback. This direct feedback of knowledge of correct response helps them to know which topics require further attention and additional practice.

Within the e-learning exercises, students mostly needed to calculate results but also had

[3] The university uses the open-source online learning management system ILIAS.

Figure 6.1 – Study design


Expectancy-Value Theory
Achievement Goals
Gender & Age
Study Information
Big Five
Ability
Present-Bias Preferences
Face-to-Face Tutorial
Subjective Goals
Covariates
Who practices more?
Practice
Does practice make perfect?
Exam
Who performs better in the exam?
Measured covariates in the first week of the semester
Voluntary practice during 12 weeks in the semester
End of the semester (15th week)

Note: The figure shows the general construct of the groups of variables I collected and the timing during the semester. It can be seen in the descriptive statistics Table 6.3 which variables are included in the different groups named tin the left side of the figure.

to answer some multiple-choice questions. I did not include open text questions because of the missing time to correct them.

The e-learning exercises were uploaded weekly, but students decided for themselves if and at what time they solved an e-learning exercise. More specifically, in the fifth week, students could solve the e-learning exercises from the first and second week, or just starting with the exercises of the first week without solving any beforehand. Another possibility was that students just solved some e-learning exercises only a few days before the exam. Students were further allowed to retake the test as often as they wanted to get even better or refresh their memory right before the exam.

Each e-learning exercise had five different versions, i.e., students who repeated exercises did not necessarily get the very same exercise. Thus, if students retook the e-learning exercise, the general frame of the exercise was the same, while the exact numbers (and thereby the solutions) might have been different. I chose this setting so that students who practiced did not just get that one exercise right by knowing the results by heart.

Participating in the e-learning exercises had no additional external reward. Still, students were able to see how well they performed in each exercise and thus could get a better feeling of how well they were prepared for the exam. For example, if a student decided to take the tests every week and had 80% of them correct each week, the likelihood was high for the student to at least pass the exam.

The duration students can work on each exercise was limited by a timer. Thereby, I wanted to ensure that students focus on the exercises and distract themselves less. Additionally, this timer also resembled the setting of the exam. Students had the double amount of time compared to the problems in the exam, so they still had enough time to solve the problems.

The official exam took place at the end of the semester. This exam was divided into a first and second trial, while the first one was called *main trial*. The first trial took place one week after the end of the lecture, and the second trial would have been one week before the new semester starts. Due to the pandemic in 2020, the second trial was postponed by several weeks into the next semester. Because of that unique situation, I do not include it in the analysis. If students had not passed the first trial, students would have been allowed to write the second trial. If they had missed or had not passed that one as well, they would have had to wait for another year. However, students can also self-selected themselves into the second trial right away.

6.4 Data

The data were collected at the University of Tübingen in the course Statistics 2 for sociology students in 2019. The data are restricted to students who took the first trial exam at the end of the semester. Thereby, I lose 21 individuals who filled out the survey or participated in the e-learning environment but did not take this trial. Moreover, nine students did not participate in the survey.

We collected the survey information within the first week with an online survey (see

Figure 6.1). The survey includes measures of the expectancy-value-theory, big five (personality traits), achievement goals, present bias preferences, subjective goals, and other demographic control variables. Then, when students had worked on the e-learning exercises at the online-study website ILIAS, the data were saved. When students solved an exercise again, only the best solution was saved. Within the face-to-face tutorial, as mentioned before, I asked students whether and how well they prepared the problem sheet. Lastly, I added the exam points of the students to the dataset to see how all those variables relate to the students' achievement in this course.

About 80 students had registered for the exam, but only 67 showed up to write the exam. 53 of these students answered the survey (at least partly), which is summarized by Table 6.2.

Table 6.2 – General sample information

Specific group information of (sub)sample	Number of observations
Registered for the exam (R)	80
Attended the exam (A)	67
At least one question of the survey answered (S)	83
Worked on at least one e-learning exercise (E)	51
A ∩ S	55
A ∩ E	47

Note: Students who did not work on the ILIAS-exercises can still be used for the regressions. I recoded non-participation to zero.

Table 6.3 shows the variables for three different sample sets. The first one shows the full sample with different number of observations per variable due to missing information of the students on some of these variables. The second includes only the students for which full information is given. The third set is made of students with at least one absent variable. Comparing the second and third samples helps to see whether the dropped individuals are different from the others. Looking at the column (1), the number of observations, it is apparent, for example, that from the 55 students who answered the survey, I could calculate the performance approach variable for only 50 students. When dropping all students with at least one missing entry, I am left with 46 students.

Table 6.3 – Descriptive statistics

	Full sample			Complete obs.			Incomplete obs.		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Points in end exam (outcome)									
Points in end exam	67	51.62	19.56	46	53.08	20.03	21	48.43	18.55
Practice variables									
Practice participation	67	5.29	4.77	46	5.40	4.82	21	5.04	4.78
Mean points per practice	67	43.74	31.95	46	45.33	31.85	21	40.25	32.68
Number of trials per practice	67	1.16	0.84	46	1.08	0.71	21	1.35	1.08
Days between first trial and final exam	67	22.51	35.82	46	24.38	39.06	21	18.41	27.83
Face-to-face tutorial preparation	67	1.46	0.86	46	1.64	0.78	21	1.05	0.90
Missing dates face-to-face tutorial	67	3.55	3.81	46	2.61	2.82	21	5.62	4.83
Individual characteristics									
Female	55	0.58	0.50	46	0.54	0.50	9	0.78	0.44
Number of semester	55	4.24	2.05	46	4.35	2.15	9	3.67	1.41
Retaking Statistics 2	55	0.15	0.36	46	0.15	0.36	9	0.11	0.33
High school GPA	55	2.69	0.64	46	2.60	0.63	9	3.16	0.45
Standardized points in Statistics 1	67	0.16	0.77	46	0.32	0.73	21	-0.18	0.77
Exam in Statistics 1 written	67	0.93	0.26	46	0.93	0.25	21	0.90	0.30
Expectancy value theory									
Self-concept	55	2.50	0.56	46	2.40	0.40	9	3.06	0.90
Intrinsic value/Dispositional Interest	55	2.56	0.84	46	2.46	0.76	9	3.06	1.07
Attainment value	55	2.48	0.51	46	2.42	0.38	9	2.81	0.87
Utility value	55	3.45	0.82	46	3.38	0.81	9	3.83	0.79
Cost	55	2.15	0.70	46	2.09	0.62	9	2.44	1.03
Big five									
Conscientiousness	53	1.83	1.14	46	1.80	1.18	7	2.05	0.85
Extraversion	54	2.10	1.21	46	2.09	1.22	8	2.17	1.26
Agreeableness	53	3.05	0.92	46	3.06	0.98	7	3.00	0.38
Openness	53	5.17	0.97	46	5.25	1.01	7	4.67	0.51
Neuroticism	53	1.45	1.18	46	1.38	1.15	7	1.90	1.38
Present bias preferences									
Risk	54	0.69	0.18	46	0.70	0.18	8	0.66	0.17
Discount factor	52	0.93	0.24	46	0.94	0.25	6	0.84	0.17
Present bias	52	1.15	0.66	46	1.16	0.70	6	1.08	0.25
Achievement goals									
Mastery approach	53	5.69	0.99	46	5.64	1.01	7	5.95	0.89
Mastery avoidance	52	4.92	1.39	46	4.99	1.29	6	4.33	2.04
Performance approach	50	3.96	1.64	46	3.96	1.67	4	3.92	1.52
Performance avoidance	51	3.65	1.82	46	3.64	1.79	5	3.73	2.35
Subjective subject goals									
How many e-learning exercises	55	7.51	3.92	46	7.48	3.99	9	7.67	3.77
How good in the e-learning exerccises?	55	0.72	0.20	46	0.72	0.18	9	0.73	0.30

Solving the e-learning exercises weekly?	55	1.45	0.66	46	1.50	0.69	9	1.22	0.44
Which grade in exam?	55	2.26	0.65	46	2.30	0.69	9	2.09	0.40

Note: The table shows the number of observations, the mean and the standard deviation per variable for three different sets of samples: first the raw sample in which I include all individuals who wrote the exam. The number of observations changes because some students did not answer the survey or did not answer some specific questions of the survey. Next, I look at the complete-cases sample. There, I only include individuals for which I have all variables answered. Thereby, the number of observations is fixed for this sample for all variables. Lastly, I include the sample of incomplete-cases to show whether my sample might differ due to the drop of individuals.

For the analysis, I rely on the number of points in the end exam because of the more precise variation compared to the final grades. The maximum number of points in the exam was 90, the best student achieved 87 points, and the passing cut-off was 40 points. *Practice participation* (referred to solely as *participation* from now on) is the main practice variable, showing the sum of e-learning sessions students participated in. I have students who never and always participated within the data, while the mean is between five and six sessions. For the performance, I look at the mean performance of the sessions the students participated in (*mean points per practice*, also called *performance* in the text). In *number of trials (of any session) per practice*, I measure the mean number of trials per session (similar to the performance). With the help of timestamps in ILIAS, I further include the *days between first trial and final exam. Face-to-face tutorial preparation* is the mean exercise preparation (between zero and four) over the thirteen tutorial weeks. The higher the value, the more often students worked on the exercise sheets before going to the mandatory face-to-face tutorials. Then the show-up rate (*Missing dates face-to-face tutorial*) measures the number of tutorials students missed with 2 to 3 missings on average.

Slightly above half of the students in the course are females. As the course was designed for third-semester students, the mean being slightly above four indicates that I had some students who either retook the exam or just postponed it to a later semester. This is also shown in the next row for the dummy *retaking Statistics 2* with a mean slightly above zero. The mean high school GPA is about 2.6 (in Germany, the HS GPA goes from 1, best, to 4, worst). Here, the incomplete sample is a bit worse, with a mean of 3.16. This also the case for the subject specific ability measure *Social Science Statistics I*(called only Statistics 1 from now on). For this variable, I standardized the number of points for the specific exam date to make the achievement comparable. Further, I included a variable indicating an individual had not passed the exam yet and if she had not written the exam at all.

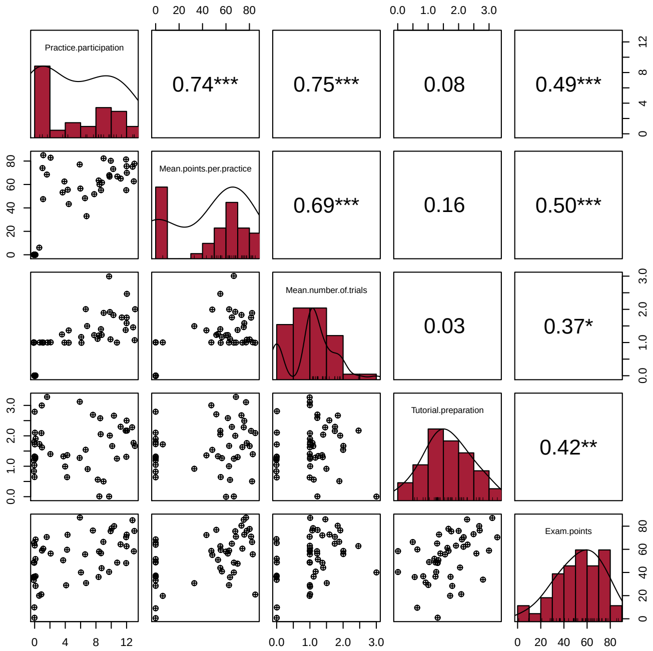
Comparing the means of the named variables, they reveal little meaningful differences. Notable exceptions are the following: individuals with at least one absent information are about 5 (out of 90) points weaker in the end exam and performed a bit worse in the e-learning exercises. Additionally, students in the complete case subsample missed, on average, only 2.6 face-to-face tutorials, while in the incomplete case subsample students missed, on average, 5.6 dates. Retakers should especially drive this difference. They were not obligated to come to the tutorials anymore and, thus, were less likely to answer the survey questions. Aside from the lowered attendance, they were also slightly less prepared for the tutorial. Then, the incomplete sample is a bit worse in the high school GPA, with a mean of 3.16 compared to 2.6. This seemingly weaker ability is also the case for the exam performance in Statistics 1.

Given the free choice, selection into practice is highly likely. Due to Germany's ethical standards, there is no possibility to use classical randomization of students excluding some students from the e-learning exercises. Therefore, practice variables might also capture the ability or motivation of the students. To address this problem, I surveyed essential variables explaining exam grades. The general list was already included in Figure 6.1 and is more specified in Table 6.3. Adding to the ability measures mentioned above, I surveyed standard measures of the expectancy-value theory (source: Gaspard et al., 2017, adapted to the university context and course), achievement goals (source: Elliot & Murayama, 2008, translated and adapted for the specific context), the big five personality traits (source Schupp & Gerlitz, 2014, taken as is) and present bias preferences (source: Frederick et al., 2002, translated). The respective variables do not show anything extraordinary for the sample. In this chapter, the variables serve only as control variables to reduce a possible omitted variable bias and are not the main focus and thus not explained more in-depth.

In addition to the variables mentioned above, I asked students' subjective goals for the practice and the exam. I asked (i) how many of the e-learning exercises they planned to solve and if they did the additional e-learning exercises, whether (ii) they planned to take them weekly as well as (iii) how well students wanted to perform in them and (iv) which grade they want to achieve in the final exam. Students wanted to complete seven to eight e-learning exercises at the beginning of the semester. Lastly, on average, students aimed for a 2.3 at the exam.

Differences between participation, performance, the number of trials, and preparation can only be measured if these variables are not multicollinear. Hence, Figure 6.2 illustrates the relationship between the four practice variables and exam grade.

Figure 6.2 – Correlation plot



Note: The diagonal shows the distribution of the respective one-dimensional distribution. The lower half shows the two-dimensional scatterplot and the upper half the correlation. $^{+}\ p < 0.10$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

We can see that the correlation between participation and performance is fairly high (0.74). The scatter plot for these variables reveals that one reason for the high correlation is the students who never participated in the e-learning sessions. Then, there are also no students with a high number of sessions and a very bad performance in these sessions. This is similar but less extreme for the number of trials. The correlation of the face-to-face tutorial preparation is very low and insignificant for the participation (0.08) and the performance (0.16). All of the practice variables are positively correlated with the exam points. In the appendix, I further included a graphical overview of the performance of the students for the weekly e-learning exercises. They show a high variation of the performance even for the students with a good grade at the end.

6.5 Model

The relationship between practice and exam points is estimated using an OLS regression:

$$\begin{aligned} points_i = \alpha &+ \rho_1 \; participation_i + \rho_2 \; performance_i + \rho_3 \; number\ of\ trials_i \\ &+ \rho_4 \; time\ between\ first\ exercise\ and\ exam_i + \rho_5 \; preparation_i \\ &+ \rho_6 \; face\text{-}to\text{-}face\ tutorial\ show\text{-}up\ rate_i + \epsilon_i, \end{aligned} \tag{6.1}$$

where the index i stands for the individuals and ϵ_i is the idiosyncratic error term. The outcome variable $points_i$ is the number of points in the end exam. For the estimation of the students' practice behavior, the equation includes at first the variable $participation_i$ to measure the effect of the additional practice solely through the e-learning exercises on exam grades. Next, $performance_i$ is the mean share of correctly answered questions per session students participated in. The $number\ of\ trials_i$ measures whether students retook sessions or not. It is, again, the mean over the weekly sessions they worked on. The $time\ between\ the\ first\ exercise\ and\ exam_i$ is in days and captures whether students worked during the semester or just at the very end. $Preparation_i$ is the self-reported preparation of face-to-face tutorial and, thus, reveals to what extent students practiced aside from the e-learning exercises. The face-to-face tutorial show-up rate is another measure for the students' tutorial participation. Students usually missed only one or two sessions if at all, because only retakers were allowed to miss more often.

The practice variables may be influenced by confounders, like motivation, personality traits or achievement goals and alike. For example, motivation might lead to an increase in additional practice. More motivated students practice more often and perform better in the e-learning sessions. Thus, the variables might not only measure the practice effect but also include the underlying motivation. Therefore, I would like to run an additional regression with necessary practice variables and additional control variables like the following:

$$\begin{aligned} points_i = \mu &+ \boldsymbol{\rho}' \, \boldsymbol{practice}_i \\ &+ \boldsymbol{\beta}_1' \, \boldsymbol{char}_i + \boldsymbol{\beta}_2' \, \boldsymbol{EVT}_i + \boldsymbol{\beta}_4' \, \boldsymbol{achievement\ goals}_i \\ &+ \boldsymbol{\beta}_4' \, \boldsymbol{big\ five}_i + \boldsymbol{\beta}_5' \, \boldsymbol{subjective\ goals}_i + \boldsymbol{\beta}_6' \, \mathbf{X}_i + \eta_i, \end{aligned} \tag{6.2}$$

where the index i stands for the individuals, and η_i is the idiosyncratic error term. $\boldsymbol{practice}_i$, $\boldsymbol{ability}_i$, $\boldsymbol{EVT}_i$, $\boldsymbol{achievement\ goals}_i$, $\boldsymbol{big\ five}_i$, $\boldsymbol{present\ bias\ preferences}_i$, $\boldsymbol{subjective\ goals}_i$ and $\boldsymbol{X}_i$ are each vectors of variables. The vector of variable $\boldsymbol{practice}_i$ includes a set of the above mentioned practice variables, which revealed to be of importance in the equation beforehand. With the additional control variables, I want to reduce the possibility of a biased estimation of the practice variables. For that, I include the high school GPA, and the Statistics 1 grade in $\boldsymbol{ability}_i$. Then, to measure motivation, I include items of the expectancy-value-theory ($\boldsymbol{EVT}_i$) and achievement goals ($\boldsymbol{achievement\ goals}_i$). I control for different types of personalities of the students due to the big five personality traits in $\boldsymbol{big\ five}_i$. Additionally, I asked the students present-bias-preferences ($\boldsymbol{present\ bias\ preferences}_i$) and their goals for this semester ($\boldsymbol{subjective\ goals}_i$). Lastly, I include additional demographic covariates in $\mathbf{X}_i$ such as gender, age, and the number of semesters. The complete list of variables is presented in Section 6.4, Table 6.3.

However, estimating the regression is problematic because of the small sample. Hence, I use variable (or feature) selection methods before running any regression with additional variables. Therefore, I follow the double selection procedure introduced in Belloni et al. (2013). Their suggestion is a two-stage selection procedure: first, variables are selected to explain (a) the outcome variable and (b) all regressors of interest, here, the practice variables. Thereby, I include variables that are not only important for the exam but also for practice behavior of the students. Second, I run an OLS regression, including the pre-selected variables and the variables of interest on the outcome. Assuming that the most important variables were surveyed in the first place, I interpret the estimated practice coefficients cautiously causal. For the feature selection, I follow Belloni et al. (2013) and use the Lasso to get sparse models for the exam points as well as the practice variables. This procedure is also applied in Urminsky, Hansen and Chernozhukov (2016). Then, I further use the elastic net as well as the Random Forests instead of the Lasso to use another algorithm for the double selection process to look for robustness.[4]

6.6 Results

Table 6.4 shows the results for the practice variables without any additional control variables. The first column includes only the variable *participation* to show whether participation in the additional online exercises predicts more points in the end exam. The coefficient is equal to 1.917 and highly significant. Thus, students who practiced one additional e-learning session increased their points on average by around 2 points. Since there were 13 sessions, students with complete participation improved their grades by more than one full grade. Next, I include the *performance* of the e-learning exercises and the *preparation* for the face-to-face tutorials. The inclusion leads to a slight decrease of the *participation*-coefficient down to 1.247, while the coefficient for the *performance* is equal to 0.089, and the one for the preparation is equal to 6.463. Only participation and preparation are significant, meaning that there is not necessarily an additional effect of the performance in the e-learning exercises on exam points. The estimated practice

[4] More specifically, I used the R-packages `glmnet` written by Friedman et al. (2001) for the Lasso, `randomForest` developed by Liaw and Wiener (2002) and `Boruta` written by Kursa and Rudnicki (2010) for the Random Forest.

effect of the e-learning participation effect is lower than for the preparation of the face-to-face tutorial. This can be explained by the voluntariness of the e-learning exercises in contrast to the obligation to attend the tutorial. In the face-to-face tutorial, the exercises were explained to the students, which might have a bigger effect when the tutorial was properly prepared. The e-learning exercises only include the additional self-testing of students.

Table 6.4 – Main practice regression - sequential inclusion of practice variables

	Dependent variable: Points in end exam					
	(1)	(2)	(3)	(4)	(5)	(6)
Practice participation	1.917***	1.247*	1.499**	1.238*	1.193*	1.539**
	(0.430)	(0.712)	(0.763)	(0.720)	(0.688)	(0.745)
Mean points per practice		0.089	0.117	0.092	0.084	0.130
		(0.120)	(0.124)	(0.122)	(0.121)	(0.130)
Mean tutorial preparation		6.463***	6.115***	6.948**	5.945**	6.123**
		(2.176)	(2.164)	(2.775)	(2.441)	(3.023)
Number of trials per practice			−3.011			−4.578
			(3.163)			(3.351)
Missing dates face-to-face tutorial				0.192		0.359
				(0.659)		(0.701)
Days between first trial and exam					0.041	0.057
					(0.052)	(0.048)
Constant	41.485***	31.717***	33.164***	30.262***	32.040***	31.636***
	(3.511)	(4.513)	(4.956)	(7.057)	(4.588)	(7.251)
Adjusted R^2	0.207	0.282	0.278	0.271	0.275	0.266
Observations	67	67	67	67	67	67

Note: Heteroskedastic robust standard errors in parentheses. $^+$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Adding the mean of the number of trials of the participated e-learning sessions in column (3) does not substantively change the regression, and the coefficient itself is statistically insignificant. The estimation is further robust to include the tutorials' show-up rate and the time range from the first exercise to the exam. The last column adds all practice variables into the regressions with no additional insight. Further, the adjusted R^2 peaks in column (2), giving evidence that the inclusion of the additional practice variables is no improvement to the model.

The estimated coefficients above should be interpreted cautiously because they could be biased due to omitted variables. Therefore, I add additional control variables in the next section.

6.6.1 Lasso, elastic net and random forest feature selection

Since the data contains almost as many variables as observations, there might not be enough degrees of freedom for meaningful inference, including all variables. To select the variables, I follow Belloni et al. (2013), selecting the most important variables for the outcome and additionally for the practice variables. The union of all selection rounds is used for a post-selection OLS regression. Thereby, I should have a high explanation for the outcome and reduced potential biases of the practice variables. To select the variables, Belloni et al. (2013) advice to use the Lasso. I further include the elastic net with α equal to .6 and .4 because the Lasso might fail if the correlation of regressors is too high. Additionally, I use the random forest algorithm to see whether the results depend on the algorithm in use. Table 6.5 panel A presents the final post-selection OLS regressions for all available individuals. To compare the results again with the complete-case sample, I include the estimates of the latter in panel B in Table 6.5.

For all four post-double selection regressions, the estimated coefficients of the participation and the preparation are similar to before, and the performance coefficient is still statistically insignificant. Panel B illustrates the complete-case sample in which the estimated participation effect is not statistically significant in two columns. However, one should note that the estimated coefficient is robust, and only the standard errors increase with the slight decrease of observations. Therefore, I would still interpret the *participation* coefficient as meaningful.

Thus, even after controlling for a rich set of covariates, there is still an effect of participation and preparation on the exam points. As mentioned earlier, I have no randomization, allowing only some students access to the e-learning exercises. Nevertheless, even after including a very rich set of control variables, there is still a practice effect. I.e., I can neglect that the practice effect is purely driven by motivation, goals, personality, or ability.[5]

[5] In the appendix, I additionally included regression specification of Equation (6.2). Given the low degrees of freedom, standard errors should be seen with caution, but the coefficient estimates are very robust.

Table 6.5 – Post-double selection regression results

	Dependent variable: Points of the end exam			
	Lasso (1)	Elastic net (.6) (2)	Elastic net (.4) (3)	Random forest (4)
Panel A				
Practice participation	1.333*	1.528*	1.247*	1.865**
	(0.731)	(0.833)	(0.712)	(0.747)
Mean points per practice	−0.064	−0.061	0.089	−0.161
	(0.122)	(0.138)	(0.120)	(0.136)
Face-to-face tutorial preparation	7.255***	6.473***	6.463***	7.948***
	(2.476)	(2.226)	(2.176)	(2.570)
Adjusted R^2	0.585	0.537	0.282	0.599
Observations	55	53	67	50
Panel B				
Practice participation	1.328	1.553*	1.316	1.445*
	(0.831)	(0.875)	(0.806)	(0.858)
Mean points per practice	−0.041	−0.060	0.127	−0.027
	(0.135)	(0.143)	(0.142)	(0.150)
Face-to-face tutorial preparation	8.367***	7.868***	9.422***	10.253***
	(2.523)	(2.375)	(2.594)	(2.580)
Adjusted R^2	0.589	0.571	0.370	0.616
Observations	46	46	46	46
Number of selected controls	3	3	0	15

Note: Heteroskedastic robust standard errors in parentheses. Panel A uses the individuals without missings for the selected variables. Panel B uses the complete cases sample. Regression results show the OLS results after a double-selection. Column (1) uses the Lasso for the selection, column (2) the elastic net (EN) with $\alpha = 0.6$, column (3) the EN with $\alpha = 0.4$, and column (4) the Random Forest (RF). The row with the number of control variables shows how many extra variables were seleted by the double-selection procedure for both panel A and B. The full set of estimated coefficients can be seen in Tables E1 and E2 in the appendix. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.7 Conclusion

This study analyzed a voluntary, non-rewarding e-learning environment that was offered to students to practice the lecture and tutorial subject topics for a statistics course. Therefore, I followed social sciences bachelor students' practice behavior during a whole semester and matched it to the exam points at the end.

I found a positive effect for the e-learning participation on the final exam points, which was not confounded by motivation, self-concept, nor ability and alike. The estimated coefficient for the performance in these e-learning sessions is statistically insignificant and varied its direction depending on the regression's specifications. Additionally, to account for general engagement in the course, I surveyed the weekly preparation of the face-to-face tutorial, which showed a substantial impact on the exam grades. This study shows that practicing statistics, conditional on a rich set of control variables, is helpful for the final exam. Hence, this chapter underlines the results of the previous chapter using a different setting.

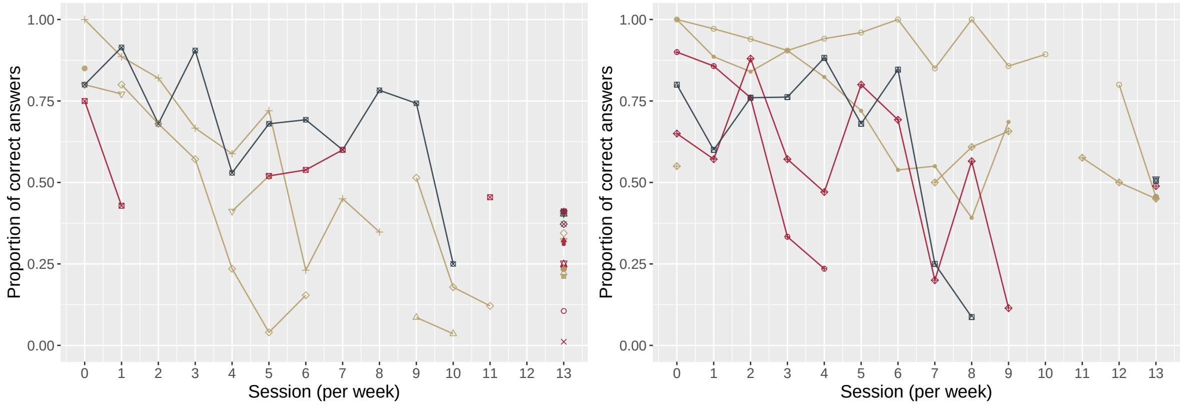
As in the previous chapter, the causal interpretation, however, is limited by two problems: first, the rather small number of individuals who gave full information and took the exam and second, the fact that I need to assume to have surveyed the necessary variables.

Appendix

E.1 Graphical overview of exercise participation

Figures E1 to E4 give an overview of the student's performance in each e-learning exercise. The first three figures sort the students by exam grades, while the last sorts students in the respective subplots by the participation rate. Since students were able to solve the e-learning exercise at any point in time during the semester, one cannot necessarily see a time trend of the students. It is possible that students re-did an e-learning exercise or solved a later one before an earlier one. Week 12 had two e-learning exercises, and exercise 14 the exam. The first set of figures groups the individuals by grades, the second by the cumulation of weeks they have a performance on. That means students who took the exam, the pretest and solved three e-learning exercises have a value of five. I see that students who did not pass the exam have either not participated in the practice or have a slightly negative trend to the end of the sessions.

Figure E1 – Performance variation, grouped by grades

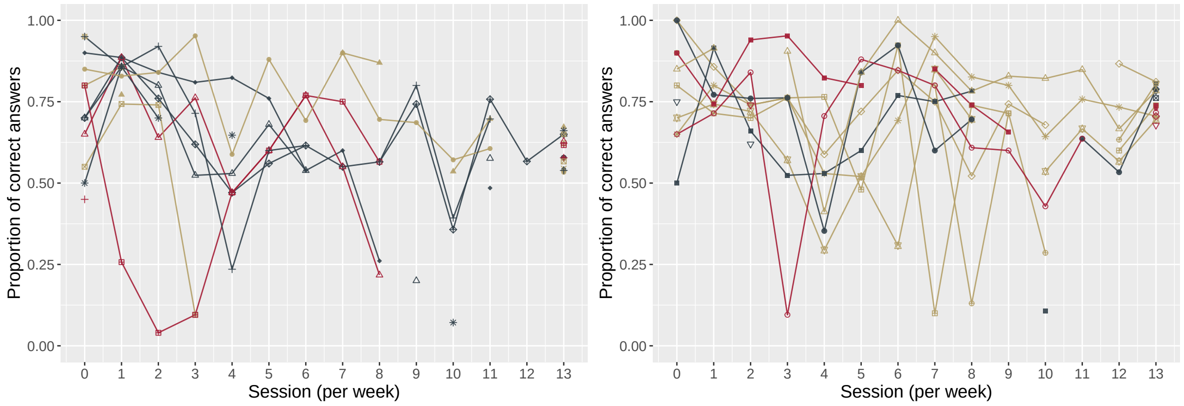


(a) Exam grade = 5 & Points > 0

(b) 3.7 ≤ Grade ≤ 4

Note: Each line in the figures shows the (best) individual results of the students for each week he or she solved an e-learning session. The students are grouped by their exam results, which is highlighted in the subcaptions. One cannot necessarily see a time trend in the subfigures, because students were allowed to multiple-try the e-learning exercises without any order. It does, however, seem, that students with worse grades also got lower results for the last e-learning exercises. I excluded individuals with no voluntary participation in the e-learning exercises. In total, there are 19 individuals related to the cutoff values of the subfigure (a), 7 in (b). Not only are the results worse for students with lower exam grades, but also fewer students are included in the graph, depicting the lower participation rate. In other words: more students with low exam results failed to participate ones.

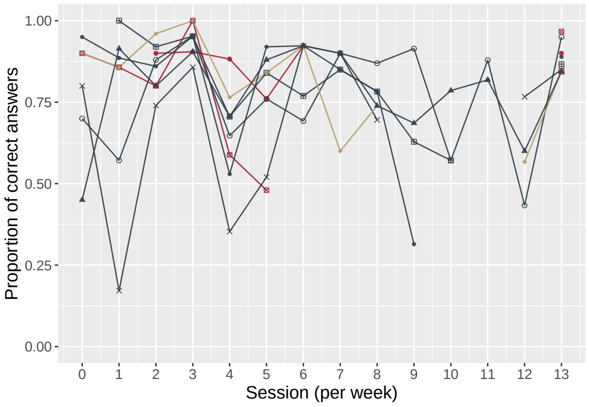
Figure E2 – Performance variation, grouped by grades



(a) $2.7 \leq \text{Grade} \leq 3.3$ **(b)** $1.7 \leq \text{Grade} \leq 2.3$

Note: Each line in the figures shows the (best) individual results of the students for each week he or she solved an e-learning session. The students are grouped by their exam results, which is highlighted in the subcaptions. One cannot necessarily see a time trend in the subfigures, because students were allowed to multiple-try the e-learning exercises without any order. It does, however, seem, that students with worse grades also got worse results for the last e-learning exercises. I excluded individuals with no voluntary participation in the e-learning exercises. In total, there are 17 individuals related to the cutoff values of the subfigure (a), 15 in (b)

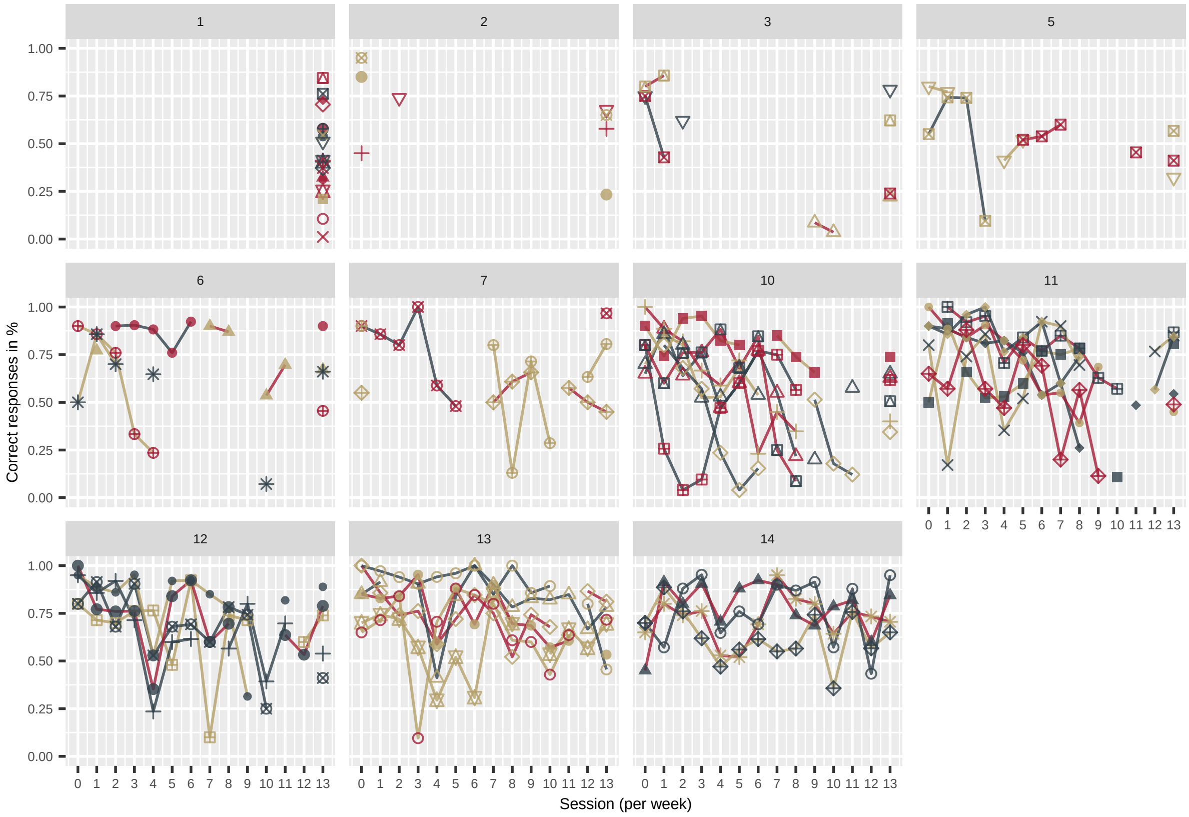
Figure E3 – Performance variation, grouped by grades



(a) $1 \leq \text{Grade} \leq 1.3$

Note: Each line in the figures shows the (best) individual results of the students for each week he or she solved an e-learning session. The students are grouped by their exam results, which is highlighted in the subcaptions. One cannot necessarily see a time trend in the subfigures, because students were allowed to multiple-try the e-learning exercises without any order. It does, however, seem, that students with worse grades also got lower results for the last e-learning exercises. I excluded individuals with no voluntary participation in the e-learning exercises. In total, there are 9 individuals related to the cutoff values in this figure.

Figure E4 – Performance variation, grouped by completed sessions



Note: The graphs groups the individuals not by grades but by the number of e-learning exercises the students participated in. There were no students with four, eight or nine participations. The students with a higher participation rate are in general better performing.

E.2 Tables

Table E1 – Full regression results of Table 6.5 panel A

	Dependent variable: Points of the end exam			
	Lasso	Elastic net (.6)	Elastic net (.4)	Random forest
	(1)	(2)	(3)	(4)
Practice participation	1.333*	1.528*	1.247*	1.865**
	(0.731)	(0.833)	(0.712)	(0.747)
Mean points per practice	−0.064	−0.061	0.089	−0.161
	(0.122)	(0.138)	(0.120)	(0.136)
Face-to-face tutorial preparation	7.255***	6.473***	6.463***	7.948***
	(2.476)	(2.226)	(2.176)	(2.570)
Standardized points in Statistics 1	15.248***	16.211***		17.674***
	(3.130)	(3.814)		(4.642)
Number of semesters	1.229*			1.956**
	(0.704)			(0.979)
Intrinsic value (EVT)	4.533**			0.638
	(1.990)			(3.380)
Female		−2.575		2.014
		(4.235)		(4.400)
Mastery approach (AG)		−2.140		−6.942***
		(1.932)		(2.516)
Retaking Statistics 2				9.079
				(7.243)
High school GPA				−0.180
				(3.737)
Exam in Statistics 1 written				8.867
				(12.319)
Mastery avoidance (AG)				−0.209
				(1.755)
Performance avoidance (AG)				−0.171
				(1.301)
Self-concept (EVT)				−0.348
				(4.801)
Utility value (EVT)				2.846
				(2.967)
Neuroticism (BF)				2.447
				(1.590)
How good in the e-learning exerccises?				21.366
				(15.765)
Which grade in exam?				−4.968
				(5.018)
Constant	16.120*	46.616***	31.717***	36.398
	(8.954)	(12.704)	(4.513)	(40.397)
Observations	55	53	67	50
Adjusted R^2	0.585	0.537	0.282	0.599

Note: The table shows the full set of estimated coefficients of Table 6.5 panel A. Heteroskedastic robust standard errors in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table E2 – Full regression results of Table 6.5 panel B

	Dependent variable: Points of the end exam			
	Lasso	Elastic net (.6)	Elastic net (.4)	Random forest
	(1)	(2)	(3)	(4)
Practice participation	1.328	1.553*	1.316	1.445*
	(0.831)	(0.875)	(0.806)	(0.858)
Mean points per practice	−0.041	−0.060	0.127	−0.027
	(0.135)	(0.143)	(0.142)	(0.150)
Face-to-face tutorial preparation	8.367***	7.868***	9.422***	10.253***
	(2.523)	(2.375)	(2.594)	(2.580)
Standardized points in Statistics 1	13.631***	15.771***		12.551***
	(3.522)	(3.991)		(4.859)
Number of semesters	1.099			1.414
	(0.779)			(1.002)
Intrinsic value (EVT)	4.432*			2.258
	(2.512)			(3.785)
Female		−3.183		0.292
		(4.255)		(4.106)
Mastery approach (AG)		−2.375		−9.134***
		(2.159)		(3.097)
Retaking Statistics 2				13.616*
				(7.150)
High school GPA				−0.962
				(4.079)
Exam in Statistics 1 written				15.332
				(14.288)
Mastery avoidance (AG)				1.739
				(2.444)
Performance avoidance (AG)				−0.279
				(1.375)
Self-concept (EVT)				−0.558
				(4.018)
Utility value (EVT)				2.655
				(3.164)
Neuroticism (BF)				0.821
				(1.663)
How good in the e-learning exerccises?				20.749
				(16.397)
Which grade in exam?				−2.846
				(5.438)
Constant	13.954	44.576***	24.709***	26.708
	(10.294)	(13.764)	(6.383)	(45.068)
Observations	46	46	46	46
Adjusted R^2	0.589	0.571	0.370	0.616

Note: The table shows the full set of estimated coefficients of Table 6.5 panel B. Heteroskedastic robust standard errors in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 7

The added value of hints in multiple-try feedback: Can feedback enhance students' achievement during the semester?[*]

[*] This chapter is based on: Schwerter, J., F. Wortha and P. Gerjets (2020): The added value of hints in multiple-try feedback: Can feedback enhance students' achievement during the semester?, unpublished manuscript, University of Tübingen.

7.1 Introduction

Feedback is one of the most powerful and extensively researched educational interventions. Multiple reviews and meta-analyzes showed that feedback interventions can have a substantial positive effect on learning and learning outcomes (e.g., Azevedo & Bernard, 1995; Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Hattie & Timperley, 2007; Jaehnig & Miller, 2007; Kluger & DeNisi, 1996; Kulhavy, 1977; Kulik & Kulik, 1988). However, these reviews have also shown that the effect of feedback varies enormously, with strong negative effects when the feedback is not designed or implemented properly (Hattie, Gan & Brooks, 2016). Potential reasons for these inconsistencies are the different approaches and paradigms that are used to investigate feedback (Narciss, 2006). Shute (2008) reviewed the literature and derived a list of feedback guidelines describing which type of formative feedback enhances learning, and which type of feedback impairs learning.

The literature found four important requirements for the effectiveness and usefulness of feedback. Shute (2008) names the following: (i) the student is in a situation in which s/he needs feedback, (ii) s/he gets the feedback in time, and (iii) s/he is able and willing to make use of the feedback. Clearly, the feedback must be appropriate for the task and the learner's predispositions, but the student must commit her/himself to use it. Following Azevedo and Bernard (1995), Kluger and DeNisi (1996) and Proske, Körndle and Narciss (2012) (iv) feedback must be further about the task itself and not evaluate the individual person to be effective.

Within computer-based learning environments, feedback plays a particularly significant role, because feedback can be provided immediately and personalized in these environments in ways that are not feasible for human tutors. Accordingly, a broad body of literature has investigated the design and effectiveness of feedback in computer-based learning environments. A recent review has shown that results are in line with the aforementioned investigations on feedback in other educational settings (Van der Kleij, Feskens & Eggen, 2015). Specifically, feedback interventions have shown a substantial positive effect on learning and learning outcomes on average, with greatly varying effectiveness.

When broken down into well-established forms of feedback, results showed that elaborative feedback (explaining why a response was correct or incorrect) was more effective than less elaborative forms of feedback (e.g., knowledge of correct response or knowledge of response). These effects were more pronounced in STEM disciplines (i.e., mathematics) as compared to the social sciences.

One form of feedback that was found to be particularly effective in fostering higher-order learning is multiple-trial feedback. Multiple-try feedback describes a feedback process in which students are informed that their answer is not correct (with varying amounts of elaboration). The learners subsequently have a chance to correct their errors. After the students have answered the question correctly or reached the maximum number of trials, knowledge of correct response is provided. A review of multiple-try feedback (Clariana & Koul, 2005) has found that the latter was more effective than other forms of feedback for higher-order learning outcomes, but inferior for lower-level outcomes. The authors argued that the generative effect this feedback encompasses is particularly high for tasks that require students to develop a deeper understanding (rather than learning facts).

Attali (2015) investigated the use of several types of feedback in mathematical problem solving: (i) multiple-try feedback with and without additional hints, (ii) knowledge of correct response, and (iii) no feedback. He analyzes this for multiple-choice and open-ended questions. In this experiment, participants had to complete 15 items and received feedback depending on the experimental condition. Subsequently, they worked on similar items but received no feedback. Results showed that multiple-try feedback leads to higher learning gains than knowledge of correct response without multiple tries and no feedback. Moreover, multiple-try feedback with hints was more effective in fostering learning than multiple-try feedback without hints. Lastly, learning gains were higher for open-ended questions when compared to multiple-choice questions. Taken together, this study showed that multiple trial feedback is effective for higher-order STEM-related learning outcomes for multiple-choice and open-ended questions, particularly when the feedback includes not just knowledge of correct response but also additional hints. The authors explain these results through the mindful and effortful problem solving that multiple trial feedback elicits. However, while the research outlined above showed that multiple-trial feedback is very promising to foster high-order learning, a transfer of these results to educational practice is still missing.

To our knowledge, there are up to now still no studies that were directly investigating the effectiveness of such feedback in university courses. Multiple-try feedback in computer-based learning environments is directly applicable to central areas of higher education, such as math or statistics education. Solving mathematical and statistical problems is a core element to many fields of studies, which requires students to go through complex chains of calculations and revise answers if necessary. This is, in particular, valid in fields of social sciences, which are less mathematical but still require students to pass several statistics courses.

Therefore, in the present study, we aim to address this research gap by investigating the additive value of hints in multiple-try feedback in an undergraduate statistics course for social sciences students in weekly e-learning tutorials. To this end, we aim to test the following hypothesis derived from the literature: does multiple-try feedback with hints outperforms multiple-try feedback without hints in a statistics class for social sciences students at a university during the whole semester. In a second step, we further analyze the performance in the e-learning sessions on the exam grade to give evidence to the general prospects of multiple-try feedback in higher education courses.

To measure the effect of the additional hint, we used the performance in twelve e-learning sessions as well as tests immediately thereafter and a one week delayed test. We differentiate between multiple-try feedback with knowledge of correct response (MTC) and multiple-try feedback with hints after an initial incorrect response (MTH). To this end, students received weekly alternating MTC or MTH in e-learning sessions, as shown in Table 7.1. We observed 87 students with varying participation within the weekly e-learning sessions.[1] The course is set within the third semester of a sociology bachelor's program at a large Germany university. The majority of the students are 20 to 23 years old, and about 60% are female. Students were informed about the data collections and gave their consent, and the experiment was ethics-approved by the faculty.

Table 7.1 – Within variation of treatment during the semester

	Session 1	Session 2	Session 3	Session 4	Session 5	...	Session 12
Group 1:	Treatment	Control	Treatment	Control	Treatment	...	Control
Group 2:	Control	Treatment	Control	Treatment	Control	...	Treatment

Note: The treatment varied weekly between the two groups. Treatment stands here for the multiple-try feedback with additional hints (MTH) while the control group is the multiple-try feedback without hints. The students are randomly selected into either group one or two at the beginning of the semester.

Based on Attali (2015) we expected significantly higher learning gains in the MTH condition in one-week follow-up measures. Our results show that students benefitted from the hints in the learning phase itself and the week thereafter. In general, students who performed better in the e-learning sessions performed better in the exam at the end of the semester. There is, however, no treatment group effect on the exam, which was to be expected, given the within-randomization during the semester.

The article proceeds as follows. Section 7.2 describes the specific setting on the chapter. Then, Section 7.3 describes the data and Section 7.4 the empirical model. Section 7.5 present the results and Section 7.6 concludes discussing the results.

[1] We do not include a group without any feedback due to ethical concerns.

7.2 Semester structure, study design and e-learning environment

7.2.1 Semester structure and course context

The present study was conducted as a part of an undergraduate level statistics course, which consists of weekly face-to-face lectures and tutorials. This course is set within the third semester of statistical education in social sciences and covers topics in statistical inference.[2]

The experiment spanned over 15 tutorial sessions. The first and the last session were reserved for the pre- and posttest. One session before the posttest was reserved for introducing the statistical software STATA. The remaining 12 weeks were mandatory e-learning sessions, in which students were asked to solve exercises (see below). At the start of the semester, students enrolled into one of six tutorial groups. These groups determined the day and time slot of the weekly tutorial.

The face-to-face lectures are unaffected by the experiment. Students got the slides a few days before the lecture took place, and there were (some) pre-recorded lecture videos from previous semesters. Then, the tutorial was changed compared to the previous semester. The original tutorial was a classic face-to-face tutorial in which tutors explained the solution of problem sheets. In the new setting, the tutorial took place in computer labs and consisted of new e-learning sessions in which students had to solve problems (i.e., exercises) themselves. With slight (numerical) changes, these problems were a subset of the pre-uploaded problem sheets of the previous semesters. Thus, students could prepare themselves for the e-learning session by working on the uploaded problem sheets. To help students, we had further pre-recorded solution videos for each

[2] Topics: Probability theory, random variables, discrete and continuous distributions, specific distributions, multidimensional random variables, limit theorems and sampling, point estimation, confidence interval estimation, statistics test, and regression analysis.

problem sheet, because students might have felt overwhelmed by solving the problems on their own with only the help of the lecture slides.

Participating in the e-learning sessions was mandatory to be allowed to write the exam at the end of the semester. Students were allowed to miss an e-learning session not more than twice. There were two dates for the exam: one week after the last lecture and one week before the next semester started. Students were allowed to select themselves into one of the two options, but the majority opted for the first date.

All materials, as well as the e-learning sessions, were integrated into an online management system of the university.[3] The learning environment featured an overview page, which could be assessed by their student ID login. On this page, students were provided with the available contents. These included the videos, which can be viewed at any time after they have been uploaded. Students only had access of the e-learning sessions within their respective tutorial time slots.

7.2.2 Learning phase

Research has repeatedly shown that feedback conditions, in general, are significantly outperformed by any form of adequately designed feedback (for example, Attali, 2015). Given these considerations, two types of feedback were selected as an experimental manipulation: multiple-try feedback with knowledge of the correct response (MTC) and multiple-try feedback with hints after an initial incorrect response (MTH). In both conditions, participants had up to three attempts to answer the open-ended or some multiple-choice questions. After an incorrect response in the MTC condition, participants were informed that their answer was not correct and asked to try again. After three attempts the correct answer was displayed. In the MTH condition, after an incorrect response, participants were informed that the answer was incorrect, a hint (explained further below) appeared and they were asked to try again. For each question, after the initial incorrect response, no additional hint was displayed. After the last attempt, the correct

[3] The university uses the open-source online learning management system ILIAS.

answer was displayed for both groups. In case the students under the MTH-regime responded correctly, the students still got the additional hint to make sure that students in the treatment group all got the treatment.

The experimental manipulation was randomized at the individual level. Specifically, while half of the students started with a control session and subsequently alternated between experimental and control sessions, the other students started with an experimental session and alternated in the opposing order, as already shown in Table 7.1 in the introduction.[4] Due to a technical issue the permutations were not correctly switched in the second session leading to half of the participants having five experimental and seven control sessions, and vice versa. To ensure equal chances after the posttests were passed, all experimental materials were made available to all students. Thus, all students had access to the elaborate feedback messages for all sessions.

7.2.3 Structure of the e-learning session

The experimental sessions followed the same structure throughout the semester which consisted of three major parts: (1) exam questions covering the topics of the previous week, (2) the e-learning phase with the experimental manipulation and (3) an exemplary exam question covering the topics of the e-learning phase. More specifically:

(1) Students had 15 minutes to answer an exemplary (old) exam question, which was identical to the exemplary exam problems of the previous session. The exam questions were either one or two questions with corresponding sub-questions. In the first session this question covered contents of the pretest. These delayed tests, i.e. the exam questions of the previous week, consisted of multiple open-ended statistical problems. Students did not receive feedback while working on these questions. Furthermore, to mimic traditional pen-and-paper testing situations as closely as possible, these questions were displayed on a single page (with the option

[4] The lab in which students worked during the e-learning session was big enough so that students could not look at someone else's screen without considerable effort.

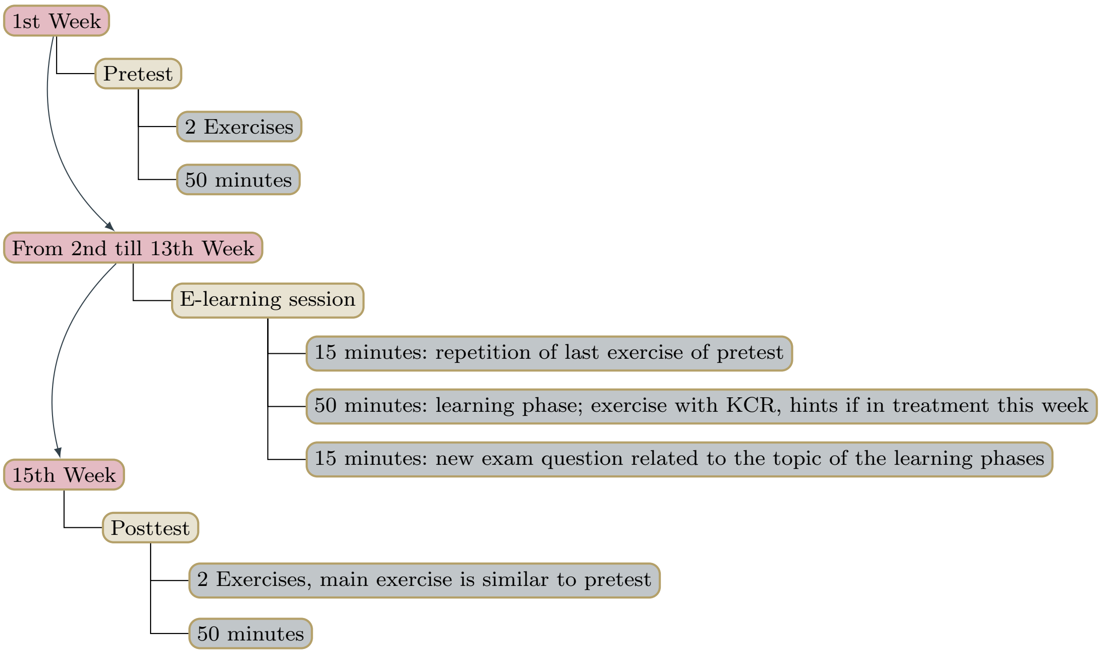
to scroll). After the 15 minutes had elapsed, students were automatically moved to the learning phase.

(2) The learning phase included the experimental manipulation. For up to 50 minutes participants worked on statistical problems. During this part each sub-problem was displayed on a single page. Here, students were required to submit an answer to proceed. When an answer was submitted, students were provided with multiple-try feedback. Specifically, when the answer was correct, students are informed that their answer was correct (i.e. knowledge of correct response), and the button that allowed them to move on became activated. When the answer was incorrect, students were informed so and told to try again. In the experimental condition an additional elaborative hint was displayed. When the third attempt was still not correct, knowledge of correct response was provided, and students were able to move to the next question. Moreover, when students provided an inadequate answer format (e.g., a word when a number was required), they received feedback that the answer was invalid, but the number of tries did not decrease. After 50 minutes, when students had not finished with this section yet, they were automatically advanced to the last phase.

(3) In the final part of the session, students had 15 minutes to work on an exemplary exam question that covered the learning phase contents. The structure of these questions was identical to the first part of the session. These problems were then the same problems at the beginning of the following week's tutorial.

Figure 7.1 provides a graphical overview of the structure. Most of the questions were of numerical nature: students had to calculate, for example, a probability or variance. In case questions hinted at the general understanding, we used multiple-choice questions in which students needed to choose an explanation or distribution. Here, students had four choices to choose from. At the very beginning and end of each session, we further asked students about self-reported emotions, which are not analyzed in depth in this chapter. The pre- and post-test sessions included self-report surveys (e.g., general ability questions), but we cannot use these because of the high number of missing values.

Figure 7.1 – Overview of the e-learning sessions


1st Week
Pretest
2 Exercises
50 minutes
From 2nd till 13th Week
E-learning session
15 minutes: repetition of last exercise of pretest
50 minutes: learning phase; exercise with KCR, hints if in treatment this week
15 minutes: new exam question related to the topic of the learning phases
15th Week
Posttest
2 Exercises, main exercise is similar to pretest
50 minutes

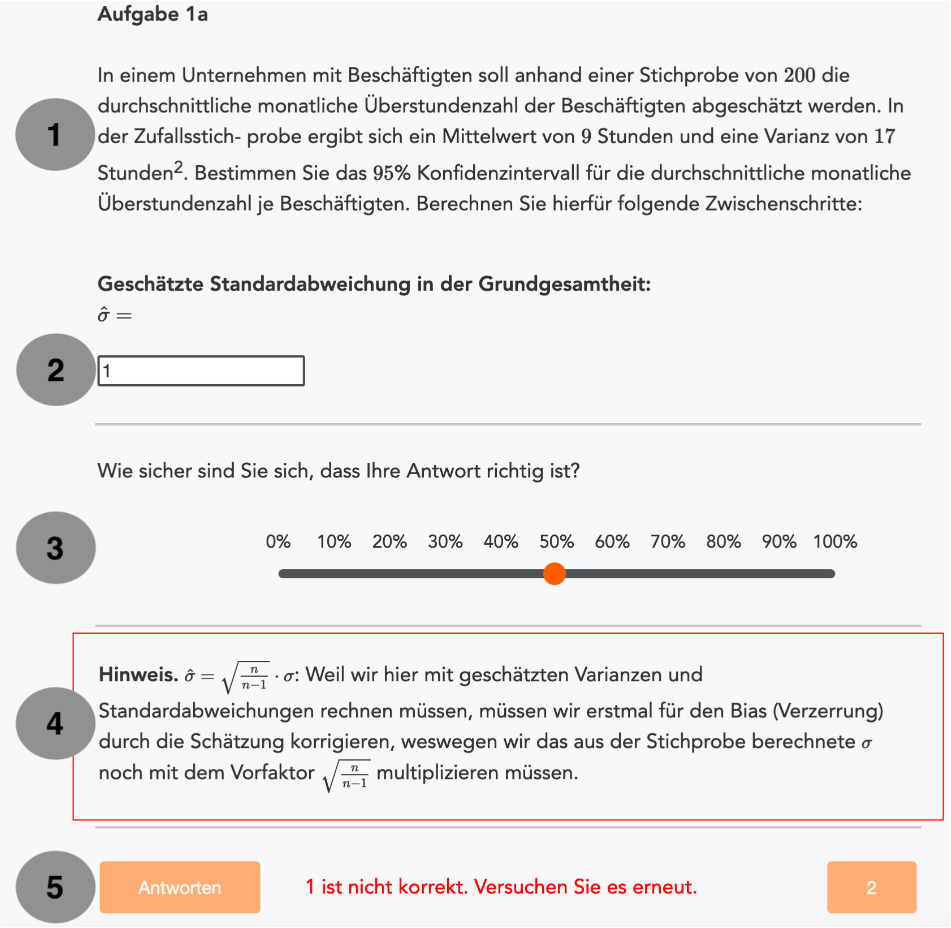
Note: The figure shows the structure of the e-learning sessions. KCR: knowledge of correct response, with or without additional hints. The exam questions within the learning environment were taken from previous years. The 14th week introduced *Stata* to the students and, thus, did not have any exercises with feedback.

7.2.4 The e-learning environment

The e-learning environment was developed to feature all contents of the tutorial sessions, including the self-reported preparation of the problem sheets and lecture participation. The e-learning environment was implemented in Qualtrics® Survey Software (Qualtrics, 2020) using JavaScript modifications and displayed in a web browser during the sessions. MathJax (Cervone, 2012) was used to display mathematical formulas. One example is shown in Figure 7.2.

The structure of the e-learning environment was linear with disabled backward navigation. Before each phase of a tutorial session, which was previously explained in the session structure, an explanation of the next phase and the corresponding time limit was displayed. The timer of the respective phase started as soon as participants advanced from this page. For the old and new exam questions, all exercises and corresponding response confidence items were displayed on a singular page. During the learning phase,

Figure 7.2 – Example of the e-learning environment



Aufgabe 1a

In einem Unternehmen mit Beschäftigten soll anhand einer Stichprobe von 200 die durchschnittliche monatliche Überstundenzahl der Beschäftigten abgeschätzt werden. In der Zufallsstich- probe ergibt sich ein Mittelwert von 9 Stunden und eine Varianz von 17 Stunden^2. Bestimmen Sie das 95% Konfidenzintervall für die durchschnittliche monatliche Überstundenzahl je Beschäftigten. Berechnen Sie hierfür folgende Zwischenschritte:

Geschätzte Standardabweichung in der Grundgesamtheit:

$\hat{\sigma} =$ 1

Wie sicher sind Sie sich, dass Ihre Antwort richtig ist?

Hinweis. $\hat{\sigma} = \sqrt{\frac{n}{n-1}} \cdot \sigma$: Weil wir hier mit geschätzten Varianzen und Standardabweichungen rechnen müssen, müssen wir erstmal für den Bias (Verzerrung) durch die Schätzung korrigieren, weswegen wir das aus der Stichprobe berechnete σ noch mit dem Vorfaktor $\sqrt{\frac{n}{n-1}}$ multiplizieren müssen.

1 ist nicht korrekt. Versuchen Sie es erneut.

Note: The figure shows an example screen of the e-learning environment. The top shows the exercises and students had to insert the answer in the white field. Below is the bar for the confidence students felt. Since the answer was incorrect in this picture, the hint in the MTH condition appeared within the red colored box.

all sub-problems were displayed on separate pages. These pages consisted of (1) the framing and question, (2) the field for the answer(s), (3) the response confidence slider, (4) the field for the elaborated feedback message, and (5) the control buttons with the field for performance feedback. When working on a question, participants were required to enter an answer in the corresponding field (2) and indicate their response confidence (4) before they were able to submit the answer using the answer button (5). Once the answer was provided, the participants received performance feedback (5) and an elaboration (4) depending on the experimental condition. When the answer was wrong, the fond and border were displayed in red. In the performance feedback message, students were told to try again and the response confidence slider was reset (i.e., centered at 50%). Furthermore, the counter on the next button, indicating the number of remaining tries was decreased. When the third trial was still incorrect, the answer button was disabled, and students were told to move to the next exercise. When the provided answer was correct, the feedback messages were highlighted in green and the next button (5) was enabled. For multiple-choice questions the procedure was identical, but participants had to select an answer and indicate their confidence in order to submit an answer. Lastly, for questions that had multiple answer fields (e.g., if students had to calculate an estimated confidence interval for a point estimate) the procedure was analog. To advance through the task, students could only submit one answer at a time (e.g., fill in the lower bound of an interval before being able to fill in the upper bound). Furthermore, the number of tries was counted for each answer field (for example, three tries for the low and three tries for the upper bound of the confidence interval). All submitted answers and corresponding response confidence ratings were logged.

7.3 Data

We observed 102 undergraduate sociology students from the bachelor of sciences *Social Sciences* at the University of Tübingen, who were enrolled in a statistics course for social sciences, during the winter semester 2018/2019. The course is set within the third-semester. From the 102 students who took the exam, ten students were retakers who did not participate in a single e-learning session and were excluded from the analysis. Further, five students were excluded from the analysis because they had not completed the exam of a precursor statistics course. This resulted in a final sample size of 87 for the analysis.

The experiment was part of the mandatory weekly tutorial of statistics for social sciences course. However, participation in this study was voluntary. Students who did not want to take part in the experiment still used the e-learning environment, but their data was not collected. All of the students who regularly attended the tutorial agreed to participate in the present study. Students were allowed to miss up to two tutorial sessions per semester without specific reasons. However, individuals who attended the class in a previous semester and missed or failed the exam ('retakers') were allowed to attend the course and the tutorial voluntarily. Students received no compensation for participation in the study. A local ethics committee approved the study.

Table 7.2 – Descriptive statistics: cross section data

	N	Mean	SD	Min	Max
Outcome					
Standardized points in end exam	87	0.07	0.99	-1.71	2.29
Treatment					
Treatment group 1 (of 2)	87	0.49	0.50	0.00	1.00
Exam information					
Second trial	87	0.36	0.48	0.00	1.00
Number of trials	87	0.11	0.32	0.00	1.00
Individual information					
Female	87	0.64	0.48	0.00	1.00
Age group below 20	87	0.28	0.45	0.00	1.00
Age group above 23	87	0.18	0.39	0.00	1.00
Pre-treatment ability measures					
Standardized Statistics 1 grade	87	-0.01	0.99	-2.52	1.40
Year Statistics 1 was written	87	2017.74	0.44	2017.00	2018.00
Points in pretest	87	8.63	7.21	0.00	24.00
Missed pretest	87	0.17	0.38	0.00	1.00
Posttest					
Points in posttest	87	7.02	6.29	0.00	18.67
Missed posttest	87	0.29	0.46	0.00	1.00
Global e-learning session information over 12 weeks					
Mean ratio of correct answers in the sessions over 12 weeks	87	0.58	0.27	0.00	0.95
Mean of missing exercises over each session	87	3.22	3.79	0.00	12.00
Mean of the number of mistakes per sessions	87	1.31	0.81	0.31	3.14
Global preparation counts over 12 weeks					
Number of lectures visited	87	5.39	4.06	0.00	12.00
Number of videos watched	87	4.64	3.93	0.00	12.00
Number of exercise sheets worked on	87	5.21	4.34	0.00	12.00
Number of exercise sheets solved	87	3.21	3.68	0.00	12.00

Note: Only the students who took the exam are included in this table. Further, if students did not participate in the pre- or posttest, we set their points to zero.

Table 7.2 shows summary statistics for the 87 students for which we obtained all necessary information as described above. We standardized the exam points of the first and second trials to include both in one regression. 31 out of 87 students wrote the second trial with 10 students who failed at the first trial and retook the exam at the second trial, i.e., wrote the exam two times within the semesters.[5] In the class, about 64% were females, 24 students were younger than 20 years and 16 were older than 23. Some students wrote the Statistics 1 grade in 2017, while the majority wrote the exam in the prior semester in 2018. 14 Students missed the pretest in which the mean was 8.63 points out of 24. Even more students missed the posttest (25). Therefore, we do not focus on pre- and posttest.

Within our e-learning environment we included three learning outcomes. The first was performance in the exercises in the learning phase itself. The second was an exemplary (old) exam question presented immediately subsequent to the learning phase. The third is the very same question presented again in the following week at the beginning of the session to analyze a one-week delayed learning outcome.

Lastly, we can also analyze the exam grade, for which we should not find a treatment effect given that the e-learning environment with all feedback hints were open to all students at the end. Then, students usually excessively study one to two weeks prior to exam and should catch up. There was no possibility to measure whether students were faster in repeating topics for which they received the treatment.

Over the twelve weeks of e-learning tutorial no one achieved 100% of the points, the highest score was 95%. The mean is with 58% per sub-problem above half of the points. Students missed, on average, 3.22 of the exercises within the sessions. Further, the students had, on average, a few more mistakes than exercises per session. Lastly, students self-reported at the beginning of the e-learning sessions how well they prepared themselves for the tutorial. The last four entries in Table 7.2 show that students joined the lecture in 60% of the cases. About half of the times they watched the tutorial videos; slightly more often did they work on the exerciser sheet before going to the tutorial but

[5] The retakers mentioned above are students who wrote the exam the year before.

only in 36% of the cases did they think that they solved it.

Table 7.3 – Descriptive statistics: panel data

	N	Mean	SD	Min	Max
Weekly e-learning session achievements					
Proportion of correct answer for all sub-problems	745	0.66	0.28	0.0	1.00
Effective Proportion of correct answers for finished sub-problems	745	0.70	0.27	0.0	1.00
Proportion of mistakes per exercises	745	1.09	0.89	0.0	6.42
Proportion of missing exercises	745	0.08	0.17	0.0	1.00
Bias-score	745	-25.01	22.33	-94.8	50.51
Proportion of correct answer in (old) exam question	745	0.28	0.31	0.0	1.00
Proportion of correct answer in (new) exam question	578	0.35	0.33	0.0	1.00
Treatment condition					
Treatment condition	745	0.50	0.50	0.0	1.00
Self-reported weekly offline preparation					
Visited the lecture	745	0.61	0.49	0.0	1.00
Watched the tutorial video	745	0.53	0.50	0.0	1.00
Worked on the exercise sheet	745	0.60	0.49	0.0	1.00
Solved the exercise sheet	745	0.36	0.48	0.0	1.00

Note: The table shows the variables over each e-learning session. The Bias-Score is the difference confidence (between 0 and 1) and correct responses (0 or 1) times 100. A negative score, therefore, shows under-confidence.

Additionally, Table 7.3 shows the weekly session information, which is partially summarized over all weeks at the end of Table 7.2. We measure the session achievement in seven different ways: (i) proportion of correct answer in the learning phase, (ii) the effective proportion of correct answers conditional on exercises finished, (iii) percentage of mistakes per number of exercises in the session, (iv) percentage of missing exercises per session, (v) the bias-score (confidence that the answer is correct minus whether the response was correct or not), (vi) the exam question at the end of the session (without any hints), and (vii) the repetition of the exam question after that. Furthermore, the tables illustrate the self-reported level of preparation per week.

7.4 Model

To evaluate the treatment-group effect on outcomes within the e-learning sessions, we use random and fixed effects models with clustered standard errors on the individual level:

$$Session_{it} = \rho \ Treatment\text{-}Group_{it} + \boldsymbol{\beta}' \ \mathbf{X}_{it} + \mu_i + \epsilon_{it} \ , \tag{7.1}$$

where index i identifies the students and t includes the time dimension of the twelve sessions. The outcome variable $Session_{it}$ is a placeholder for the seven session outcomes named before. When running the fixed effect regression, the general intercept and everything constant is included in the individual fixed effect μ_i. Thus, $\mathbf{X}_{it}$ includes only observed variables that change over the course of the sessions, like the weekly information on the preparation. Then, also ϵ_{it} includes only non-constant unobserved characteristics. When relying on the random effects model, however, $\mathbf{X}_{it}$ also includes some constant control variables that are presented in Table 7.2.A drawback of the random effects model is that we cannot cancel out constant unobservables anymore, thereby biasing the estimation. We will compare the fixed and random effects regression results to argue that this is not a problem.

For the fixed effects models we demean Equation (7.1), which cancels out μ_i, while for the random effects model, the subtracted mean is weighted with a ratio of the variance of the within- and between-variance.[6] Therefore, μ_i is not canceled out, and we need to assume that no constant unobserved variables are leading to a biased estimation. Very close regression results of the fixed and random effects models give confidence that the random-effects regression estimation results are unbiased.

Next, we use a basic OLS model with heteroskedastic robust standard errors to analyze the session performance and possible treatment-group effects on the exam points at the

[6] $\theta = 1 - \sqrt{\left(\frac{\sigma_e^2}{T \cdot \sigma_u^2 + \sigma_e^2}\right)}$, with σ_e^2 is the within variance and σ_u^2 is the between variance.

end of the semester. The model is as follows:

$$\begin{aligned} Exam_i = \alpha + \boldsymbol{\lambda}_1' \, \boldsymbol{performance}_i + \lambda_2 \, preparation_i \\ + \rho \, \textit{Treatment-Group}_i + \boldsymbol{\beta}' \; \mathbf{X}_{it} + \epsilon_i \; , \end{aligned} \tag{7.2}$$

where index i stands for the students, α for the intercept and ϵ_i is the idiosyncratic error term. The outcome $Exam_i$ is the standardized exam points of either the first or the second date of the exam. Our main focus in this regression is on the vector of variables $\boldsymbol{performance}_i$, and $preparation_i$ which measure the e-learning session achievements and pre-e-learning session behavior of the students respectively. For the performance, we use the mean percentage of correct responses during the twelve weeks, as well as the mean of the number of exercises not answered, and the mean of the number of mistakes per session over the twelve weeks. For the preparation, we use either counts for the number of lectures visited, tutorial videos watched, exercise sheet worked on, and completed as well as a sum of all four variables. The coefficient ρ of the variable *Treatment-Group* indicates, if statistically significant from zero, that one of the two groups benefitted more from the additional hints than the others. Since our treatment shifted weekly, and students had access to the hints two weeks before the exam, we do not expect to find a treatment-group effect on exam grades. We include this variable to check if our treatment had randomly adverse effects. $\mathbf{X}_{it}$ is a vector of all control variables and $\boldsymbol{\beta}$ the respective coefficients. The set of controls are presented in Table 7.2.

7.5 Regression results

7.5.1 E-Learning sessions

First, we analyzed if the additional hints within the session helped students to perform better in (i) the sessions, (ii) the exam questions after the session, and (ii) in the week thereafter. Table 7.4 shows the results for the random effects model for different session outcomes and Table F1 the respective fixed effects regressions. Since the estimation results are very robust, we focus on the more efficient random effects model. Column (1) explains the percentage of correct answer in a learning phase, column (2) uses the effective percentage of correct answers (conditional on tasks done); column (3) the number of mistakes, column (4) the number of missing tasks; column (5) the bias-score; column (6) the percentage of correct answers in the (old) exam question (without additional hints); column (7) the exam question in the following week.

Column (1) shows that being treated, i.e. getting additional hints, increased the performance by 3.4%. If we look only at sub-exercises done by the students (they could have gotten out of time and thus not solved everything) in column (2), the treatment coefficient increased up to 4.7%. Then, column (3) shows that students with additional feedback are about 14.3% less likely to make a mistake. We do not find a significant result for number of missing exercises in column (4). The coefficient when explaining the bias-score is again significant at the 10% level. The negative coefficient gives evidence for a reduced bias-score of around 2%, meaning that the treatment helps to bring their confidence in line with their actual knowledge.

Next, we analyzed the performance of the previous exam question right after the learning phase and the week after that. We do not find a statistically significant increase of correct answer right after the session (column 6), but a week later (column 7). Additionally, Tables F2 and F3 further include the session outcomes as explanatory variables for the previous exam questions within the e-learning sessions. Table F2 shows that the missing

Table 7.4 – Several session outcomes (Panel: random)

	Dependent variable:						
	LP	ELP	MIST	MISS	BIAS	NEQ	OEQ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment condition	0.037^{*} (0.015)	0.047^{**} (0.015)	-0.143^{**} (0.052)	0.009 (0.011)	-2.026^{+} (1.216)	0.011 (0.016)	0.046^{*} (0.023)
Lecture visited	0.013 (0.021)	0.019 (0.020)	-0.126^{+} (0.066)	0.003 (0.018)	3.861^{*} (1.706)	0.075^{**} (0.028)	0.125^{***} (0.024)
Video watched	0.109^{***} (0.030)	0.103^{***} (0.026)	-0.280^{**} (0.089)	-0.043^{+} (0.026)	−3.655 (2.653)	0.044 (0.035)	0.082^{*} (0.041)
Exercise sheet worked on	0.015 (0.031)	0.029 (0.030)	-0.271^{*} (0.118)	0.025 (0.028)	4.016^{+} (2.423)	0.058 (0.039)	−0.007 (0.043)
Exercise sheet solved	0.084^{***} (0.021)	0.069^{***} (0.020)	-0.208^{***} (0.062)	-0.032^{+} (0.017)	-4.592^{***} (1.385)	0.071^{*} (0.033)	0.088^{**} (0.030)
Statistics 1	0.045^{*} (0.019)	0.049^{**} (0.018)	-0.113^{*} (0.050)	0.001 (0.009)	0.398 (1.828)	0.042^{*} (0.020)	0.033^{*} (0.015)
Year of Statistics 1	-0.090^{**} (0.034)	-0.112^{***} (0.034)	0.314^{**} (0.120)	−0.014 (0.027)	1.083 (4.077)	−0.055 (0.036)	−0.066 (0.042)
Female	0.003 (0.031)	0.0002 (0.032)	0.042 (0.104)	0.012 (0.023)	-9.521^{**} (3.684)	−0.023 (0.033)	−0.014 (0.032)
Age group below 20	0.004 (0.035)	−0.001 (0.035)	0.050 (0.099)	−0.010 (0.023)	5.058 (4.786)	0.052 (0.040)	0.077^{*} (0.034)
Age group above 23	−0.022 (0.065)	−0.026 (0.060)	0.023 (0.173)	0.042 (0.036)	6.771 (4.441)	0.051 (0.045)	0.011 (0.062)
Pretest points	0.010^{***} (0.003)	0.008^{**} (0.003)	-0.026^{**} (0.008)	−0.003 (0.002)	−0.167 (0.367)	0.012^{***} (0.003)	0.014^{***} (0.003)
Missed pretest	0.059 (0.059)	0.019 (0.057)	−0.224 (0.178)	−0.044 (0.039)	3.514 (6.378)	0.126 (0.080)	0.029 (0.075)
Observations	718	718	718	718	718	718	557
R^2	0.189	0.194	0.192	0.038	0.034	0.170	0.285

Note: LP: Proportion of correct answers in the learning phase. ELP: Effective proportion of correct answers in the learning phase, i.e. only including exercises finished by the students. MIST: Number of mistakes per trial in the learning phase. MISS: Number of exercises missed to solve until the end of the learning phase. NEQ: Ratio of correct answers in the new (previous) exam questions. OEQ: Ratio of correct answers in the old (previous) exam questions. Standard errors in parentheses are clustered at the individual level and heteroskedastic robust. Fixed effect regression results are shown in the appendix with very similar results. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

treatment effect for the exam question at the end of the session in Table 7.4 column (6) is not influenced by either of the e-learning outcomes variables. The coefficient is still small and statistically insignificant. However, we find that performing better in the e-learning session predicts better performance in the exam questions. For the delayed testing, i.e., the exam question at the beginning of the session, Table F3 shows a rather stable estimation, confirming the result of Table 7.4 column (7). The different e-learning sessions outcome variables lower the treatment coefficient only slightly (columns 1, 2, 3, 6, and 7). In general, a reduction would just show that the improved performance in the session captures the main positive impact of the treatment on the old previous exam question. The reduction is, however, of small magnitude and should not be over-interpreted. The robust estimation gives evidence that there is a lasting treatment effect for the week thereafter, not completely captured by the performance in the e-learning session.

7.5.2 Exam

Next, we analyzed if the performance in the mandatory online sessions explains the exam grades. Therefore, we include the mean points over all sessions in column (1) of Table 7.5 , the mean missing tasks in column (2), and the number of mistakes in column (3). We then include all three online-session performance measures in column (4) to see which variables remain important. Column (5) includes the mean of the (self-reported) preparation variables, i.e., coming to the lecture, watching the tutorial videos prior to the sessions, working on the problem sheet, and solving the problem sheet.[7] Lastly, columns (7) and (8) add the statistics 1 grade and the year in which they wrote the exam as a subject specific ability measure.

The sessions' performance seems robustly statistically significant apart from column (5) and (7) when the general preparation was included. Seemingly, adding a measure for overall preparation for each week and the standardized statistics one points takes on some of the explanatory power of the sessions' performance. I.e., those who are generally good in statistics and put effort perform better in the sessions and are thus better in the exam. Then, however, the sessions' performance coefficient of column (5) and (7) is well within the confidence intervals of the other columns. The sample might be too small to detect significances for the preparation and performance variables.

As expected, the treatment variables' coefficient is highly insignificant, i.e., treatment-group allocation did not affect the exam grade.

[7] Adding the four named variables individually into the regression does not give additional insight and can be shown upon request.

Table 7.5 – Exam results including session information

	Dependent variable: Points of the end exam						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group 1	0.049	−0.046	0.0003	0.053	0.034	0.064	0.051
	(0.200)	(0.217)	(0.204)	(0.200)	(0.194)	(0.168)	(0.163)
Exam trial	0.376	0.082	0.339	0.383	0.460^{+}	0.455*	0.509*
	(0.267)	(0.264)	(0.266)	(0.269)	(0.271)	(0.232)	(0.240)
Number of exam trials	−0.088	0.066	−0.127	−0.087	−0.101	−0.029	−0.040
	(0.338)	(0.332)	(0.345)	(0.350)	(0.356)	(0.261)	(0.266)
Mean correct answers	1.903***			2.271^{+}	1.889	1.891^{+}	1.625
	(0.407)			(1.289)	(1.309)	(1.118)	(1.113)
Mean mistakes			−0.564***	0.112	0.132	0.263	0.272
			(0.125)	(0.372)	(0.383)	(0.316)	(0.322)
Sum of missings		−0.069**		0.005	0.027	−0.005	0.011
		(0.024)		(0.042)	(0.037)	(0.037)	(0.034)
General preparation					0.069^{+}		0.050^{+}
					(0.037)		(0.030)
Statistics 1						0.501***	0.486***
						(0.083)	(0.088)
Year of Statistics 1						−0.208	−0.197
						(0.226)	(0.222)
Trials	Both	Both	Both	Both	Both	Both	Both
Observations	87	87	87	87	87	87	87
R^2	0.244	0.073	0.203	0.245	0.274	0.446	0.461
Adjusted R^2	0.207	0.028	0.164	0.188	0.210	0.389	0.398

Note: The variable *mean correct answers* is short for the mean proportion of correct answers for all e-learning sessions. The variable *general preparation* is the simple sum of the variables *visited lectures*, *videos watched*, *exercise sheets worked on* and *exercise sheets solved.* A regression estimation with the four variables is very similar. Heteroskedastic robust standard errors in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

7.6 Conclusion

The chapter analyzed multiple-try feedback within a university statistics course. Specifically, we investigated the additional value of hints in multiple-try feedback messages on different outcomes over the course of a semester.

First, with regards to the added effect of hints in multiple-try feedback during the tutorial sessions, we found that students were significantly better in sessions with additional hints in their performance during the session and the one-week follow-up test. However, they did not show higher learning outcomes for exam questions directly after the learning session when receiving hints. At first glance, these findings seem to contradict previous studies. Attali (2015) found no effect of multiple-try feedback with or without additional hints in their initial learning phase (the equivalent to the learning phase in this chapter), but significant differences in the immediately following test.

However, Attali (2015) focused on a between subject comparison of multiple feedback-types (for two item types) in a small experimental study. Nevertheless, this chapter had a more extended learning session (about 50 minutes), and the tasks in this phase were partially interconnected. For instance, in one session, participants had to go through the different steps of a hypothesis test. Therefore, additional hints in early tasks might have had beneficial effects for the following tasks, which corresponds to the lower likelihood of making mistakes in sessions with additional hints in the multiple-try feedback.

However, with the one-week follow-up exercises, we directly addressed a shortcoming identified in previous research (Attali, 2015). Participants were significantly better in one-week follow-up exercises after they received multiple-try feedback with hints (when controlling for prior knowledge, performance on the same task in the previous week, and participation in the lecture in-between) indicates that the additional hints have a lasting effect in educationally relevant contexts. Furthermore, in line with findings that multiple-try feedback with hints is particularly effective after initial mistakes (Attali, 2015), we found that making mistakes during the learning phase was decreased in sessions with additional hints.

In this chapter, the additional guidance of a hint that fosters re-evaluation (for example, Corbett & Anderson, 2001; Lepper & Woolverton, 2002) might be particularly relevant, as the problems within a learning session were often related (e.g., going through all steps of hypothesis testing in one of the sessions). Therefore, the additional hint that pointed students towards the underlying statistical calculations/formulas was potentially useful for avoiding mistakes in future steps. In this context, the hint might have served as guided instruction (Kirschner, Sweller & Clark, 2006).

While students received information if their answer was correct in both conditions, guidance in the form of elaboration on how to solve the problem at hand was only provided in the experimental condition. A broad body of research has shown that guided learning is more effective than unguided learning from many perspectives (see Kirschner, Martens & Strijbos, 2004, for an overview). For instance, during active learning construction in computer-based learning environments, guided instruction was found to foster process and content knowledge construction (Ardac & Sezen, 2002). Solving a statistical problem required knowledge of the appropriate statistical method to apply (content knowledge) and the ability to apply this method to the current problem (process knowledge). In the present study and similar statistical tutorials in general, when students aren't able to solve such a problem, they must investigate a large number of potential mistakes they might have made (from using the wrong method/formula to simple slip-ups when using a formula). Even with immediate feedback that the answer is incorrect, students might be overwhelmed with this task, especially when they have low prior knowledge.

The additional guidance provided through the hints may have reduced cognitive load similar to worked examples when compared to constructing a solution without guidance (for example, Sweller, 1999; Sweller, van Merrienboer & Paas, 1998), which is particularly useful for low prior knowledge students. On the other hand, the initial try in multiple-try feedback leaves room to work freely on the topic without unnecessary guidance, which is beneficial for high prior knowledge students. With regard to potential misconceptions, multiple-try feedback with hints can be more beneficial as an elaboration that aims at appropriate methods/approaches to solve a statistical problem. Research has shown that such misconceptions are common in non-mathematically oriented students, even for less advanced statistical concepts (Mevarech, 1983). Feedback research has shown that feedback can be particularly effective in the revision of errors or misconceptions, especially when these errors were made with high confidence (for example, Kulhavy, 1977). The additional hints provided in this chapter may have addressed misconceptions more effectively.

Second, we found that the within-semester randomization of additional hints in multiple-try feedback showed no advantage for both groups on exam performance. This result was in line with our expectations and showed that equal chances for each student were provided despite the experimental manipulation during the tutorial sessions (e.g., students got access to all materials prior to the exam). More importantly, the performance during the e-learning sessions was a robust predictor of exam grades throughout multiple models, beyond the explanatory value of prior knowledge (i.e., performance in the previous statistics course exam). This result shows that the use of the e-learning tutorial implemented in this chapter had a substantial positive effect on students' statistics performance, even after controlling for subject-specific ability. Though we cannot claim causality, the results still give evidence that multiple-try feedback is a suitable way of providing guidance and fostering learning in STEM-related problem-based learning tasks (Attali, 2015; Clariana & Koul, 2005). However, a potential explanation is that students who generally invested more effort in learning statistics performed better in the tutorial sessions and the exam.

While no causal inferences can be drawn because of various opportunities for compensation students had before the exam, the results still caries potential for further research. Similar to our findings, previous studies have shown that (preparatory) e-learning courses do predict performance during the semester (for example, Fischer, Zhou et al., 2019). The use of this link between learning in e-learning environments and performance in corresponding classes (incl. exams) seems particularly promising for developing timely, individualized interventions. Specifically, the ability to track individuals learning and performance in real-time can be used to implement digital interventions that can foster learning processes as they unfold, potentially circumventing issues before they arise. The great potential of such interventions, such as prompts and feedback from pedagogical agents (for example, Azevedo & Bernard, 1995) has been repeatedly shown in laboratory settings, but a systematic transfer to applied educational settings (e.g., university courses) is still required.

As shown above, there are many potential functional mechanisms (incl. potential interactions of different factors) related to the beneficial effect of hints in multiple-try feedback in a university statistical course. However, the present study cannot shed light on the specific (cognitive) processes that caused these effects. Instead, we focused on investigating the transfer of previous findings (i.e. Attali, 2015) to an applied context. While this study is limited by the contextual constraints, such as the lack of a no feedback control group due to ethical concerns, it has shown that the additional value of hints in multiple-try feedback is robust enough to show a significant effect in such noisy environments. Furthermore, regarding issues in replicating results of experimental research studies like this can show the robustness of effects across contexts. Furthermore, future studies should extend this approach to other interventions and designs to close the gap between experimental research and educational practice.

In university contexts, in addition to their positive effects, e-learning environments are particularly promising because of their diminishing costs. While the design and implementation might be costly at first, the running costs are low, and the scaling possibilities are high. Ideally, such implementations should not replace human tutors but be used in conjunction. In this context, human tutors can focus on more complex problems. Future research should explore these possibilities in more depth.

Appendix

F.1 Tables

Table F1 – Several session outcomes (Panel: within)

	Dependent variable:						
	LP	ELP	MIST	MISS	BIAS	NEQ	OEQ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment condition	0.038^{*}	0.045^{**}	-0.132^{*}	0.005	-2.159^{+}	0.009	0.048^{*}
	(0.015)	(0.015)	(0.052)	(0.011)	(1.222)	(0.015)	(0.022)
Lecture visited	0.002	0.012	-0.133^{+}	0.010	3.943^{*}	0.070^{*}	0.096^{***}
	(0.021)	(0.020)	(0.070)	(0.016)	(1.747)	(0.028)	(0.029)
Video watched	0.109^{***}	0.095^{***}	-0.230^{*}	-0.066^{*}	-3.659	0.051	0.087^{+}
	(0.032)	(0.028)	(0.100)	(0.029)	(2.615)	(0.037)	(0.044)
Exercise sheet worked on	-0.001	0.014	-0.261^{*}	0.034	4.539^{+}	0.057	-0.032
	(0.032)	(0.029)	(0.123)	(0.028)	(2.461)	(0.042)	(0.050)
Exercise sheet solved	0.077^{***}	0.067^{***}	-0.232^{***}	-0.016	-4.470^{***}	0.038	0.051
	(0.021)	(0.020)	(0.064)	(0.019)	(1.353)	(0.036)	(0.036)
Observations	745	745	745	745	745	745	578
R^2	0.080	0.089	0.103	0.020	0.024	0.047	0.053

Note: LP: Proportion of correct answers in the learning phase. ELP: Effective ratio of correct answers in the learning phase, i.e. only including exercises finished by the students. MIST: Number of mistakes per trial in the learning phase. MISS: Number of exercises missed to solve until the end of the learning phase. NEQ: Ratio of correct answers in the new (previous) exam questions. OEQ: Ratio of correct answers in the old (previous) exam questions. Standard errors in parentheses are clustered at the individual level and heteroskedastic robust. $^{+}\ p < 0.10$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table F2 – New exam question and e-learning outcomes as explanatory variables

	Dependent variable: Proportion of correct answers in new exam questions						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment condition	0.001 (0.016)	0.001 (0.016)	0.005 (0.017)	0.013 (0.016)	0.012 (0.016)	0.004 (0.014)	0.005 (0.014)
Lecture visited	0.072** (0.027)	0.071** (0.027)	0.070* (0.028)	0.076** (0.027)	0.072** (0.028)	0.051^{+} (0.027)	0.053* (0.027)
Video watched	0.016 (0.034)	0.022 (0.035)	0.032 (0.033)	0.031 (0.034)	0.047 (0.036)	0.005 (0.032)	0.007 (0.033)
Exercise sheet worked on	0.053 (0.039)	0.051 (0.039)	0.047 (0.039)	0.065^{+} (0.039)	0.055 (0.039)	0.034 (0.038)	0.035 (0.037)
Exercise sheet solved	0.054^{+} (0.032)	0.060^{+} (0.033)	0.064^{+} (0.033)	0.065^{+} (0.033)	0.074* (0.033)	0.053^{+} (0.031)	0.058^{+} (0.031)
Statistics 1	0.031^{+} (0.018)	0.033^{+} (0.019)	0.038^{+} (0.019)	0.042* (0.019)	0.042* (0.021)	0.021 (0.021)	0.025 (0.021)
Year of Statistics 1	−0.030 (0.034)	−0.031 (0.035)	−0.042 (0.036)	−0.058^{+} (0.034)	−0.056 (0.036)	−0.020 (0.035)	−0.023 (0.035)
Female	−0.023 (0.030)	−0.022 (0.031)	−0.021 (0.033)	−0.020 (0.031)	−0.016 (0.034)	0.020 (0.033)	0.019 (0.033)
Age group below 20	0.052 (0.037)	0.052 (0.038)	0.054 (0.039)	0.049 (0.039)	0.048 (0.041)	0.027 (0.040)	0.030 (0.039)
Age group above 23	0.057 (0.037)	0.057 (0.039)	0.051 (0.042)	0.062 (0.041)	0.046 (0.047)	0.038 (0.039)	0.047 (0.039)
Pretest points	0.010*** (0.003)	0.011*** (0.003)	0.011*** (0.003)	0.012*** (0.003)	0.013*** (0.003)	0.008** (0.003)	0.009** (0.003)
Missed pretest	0.107 (0.075)	0.120 (0.076)	0.115 (0.077)	0.110 (0.082)	0.124 (0.078)	0.080 (0.064)	0.089 (0.068)
Mean correct answers	0.246*** (0.038)					0.399*** (0.072)	
Effective correct answers		0.200*** (0.037)					0.278*** (0.064)
Mean mistakes			−0.040*** (0.010)			−0.017^{+} (0.010)	−0.027* (0.011)
Mean missing exercises				−0.280*** (0.045)		−0.136* (0.055)	−0.311*** (0.048)
Bias-Score					0.001 (0.001)	0.004*** (0.001)	0.004*** (0.001)
Observations	718	718	718	718	718	718	718
R^2	0.235	0.209	0.187	0.209	0.172	0.277	0.263

Note: Random effects model. Standard errors in parentheses are clustered at the individual level and heteroskedastic robust. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table F3 – Old exam questions and e-learning outcomes as explanatory variables

	Dependent variable: Proportion of correct answers in old exam questions						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment condition	0.038^{+}	0.037	0.042^{+}	0.048^{*}	0.048^{*}	0.044^{*}	0.039^{+}
	(0.023)	(0.023)	(0.024)	(0.023)	(0.023)	(0.021)	(0.020)
Lecture visited	0.125^{***}	0.125^{***}	0.124^{***}	0.126^{***}	0.124^{***}	0.099^{***}	0.096^{***}
	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)	(0.021)	(0.022)
Video watched	0.062	0.064	0.075^{+}	0.077^{+}	0.082^{*}	0.058^{+}	0.044
	(0.040)	(0.041)	(0.041)	(0.040)	(0.041)	(0.033)	(0.031)
Exercise sheet worked on	-0.011	-0.013	-0.013	-0.003	-0.008	-0.038	-0.044
	(0.040)	(0.041)	(0.042)	(0.042)	(0.043)	(0.037)	(0.035)
Exercise sheet solved	0.071^{*}	0.077^{**}	0.086^{**}	0.083^{**}	0.090^{**}	0.053^{+}	0.046
	(0.030)	(0.030)	(0.029)	(0.030)	(0.031)	(0.027)	(0.029)
Statistics 1	0.026^{+}	0.025^{+}	0.030^{*}	0.034^{*}	0.033^{*}	0.022^{+}	0.013
	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)	(0.012)	(0.012)
Year of Statistics 1	-0.052	-0.048	-0.059	-0.070^{+}	-0.067	-0.049	-0.036
	(0.040)	(0.041)	(0.042)	(0.042)	(0.042)	(0.036)	(0.036)
Female	-0.011	-0.010	-0.011	-0.013	-0.010	-0.011	0.003
	(0.030)	(0.030)	(0.031)	(0.032)	(0.032)	(0.024)	(0.025)
Age group below 20	0.075^{*}	0.077^{*}	0.078^{*}	0.075^{*}	0.075^{*}	0.042	0.034
	(0.033)	(0.033)	(0.033)	(0.035)	(0.034)	(0.029)	(0.029)
Age group above 23	0.018	0.018	0.009	0.016	0.008	-0.020	-0.025
	(0.053)	(0.055)	(0.060)	(0.060)	(0.063)	(0.046)	(0.042)
Pretest points	0.012^{***}	0.012^{***}	0.013^{***}	0.014^{***}	0.014^{***}	0.008^{***}	0.007^{**}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
Missed pretest	0.032	0.035	0.025	0.031	0.028	0.016	0.020
	(0.068)	(0.068)	(0.071)	(0.079)	(0.076)	(0.051)	(0.053)
Mean correct answers	0.184^{***}						0.263^{***}
	(0.049)						(0.067)
Effective correct answers		0.167^{**}					
		(0.052)					
Mean mistakes			-0.022^{+}				0.019
			(0.012)				(0.012)
Mean missing exercises				-0.114^{+}			0.133^{+}
				(0.066)			(0.074)
Bias-Score					0.0004		0.002^{*}
					(0.001)		(0.001)
Previous exam question results						0.472^{***}	0.450^{***}
						(0.049)	(0.046)
Observations	557	557	557	557	557	557	557
R^2	0.337	0.333	0.322	0.289	0.283	0.497	0.540

Note: Random effects model. Standard errors in parentheses are clustered at the individual level and heteroskedastic robust. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 8

Dissertation Summary and Conclusion

The dissertation focuses on two different education topics: the impact of education decisions on labor market outcomes in the first part, and the research on e-learning environment possibilities in university statistics classes in the second part.

The literature finds a clear relationship between education and labor market outcomes. One of these is the increased employment probability of individuals with higher levels of education. However, for university graduates, it is unclear, whether a higher employment probability also means that students apply their higher education knowledge in a degree-related occupation. Chapter 2 finds that male engineering and computer sciences (EngComp) graduates are more likely to have a degree-related occupation compared to other males and females in general. For females with an EngComp degree, we find that the degree-related occupation advantage is either smaller or nonexistent compared to their male counterparts. Although Malamud (2011) argue that higher education should match individual skills and job requirements, there is an inevitable heterogeneity in degree-relatedness of occupations in Germany.

Following this divergence in degree-relatedness occupations, we assess whether the leaking STEM (or more specific, EngComp) pipeline could be patched if students are more exposed to math and natural sciences within their school years. Therefore, we take advantage of a reform in Baden-Württemberg, one of Germany's federal states. At the secondary schools for high-performing students ("Gymnasien") mandatory advanced math classes were introduced for all students in their last two school years. The share of students choosing advanced math classes was mostly around 20 to 30% for all other states. The results of this natural experiment suggest that males were not affected by the reform, but that women were affected negatively by the higher exposure to math. Women were less likely to graduate in the fields of mathematics or natural science. This, however, did not influence the leaking pipeline at the occupation stage: it appears that women just dropped out earlier, before obtaining the degree, than before. Given the missing positive effect of additional math exposure, future research should focus on earlier stages of schools to see if girls' interest in STEM or its subfields can be enhanced there.

Another important insight of the literature is the heterogeneity within university degrees: there are wage premia, at least for subgroups of populations, for “better” universities in the western world. Thus, after focusing on university graduates’ occupational decisions, Chapter 4 analyzes possible wage differences between university graduates of the same field of study. The literature has found on the extensive margin that more education, in general, leads to higher wages. Further, for states like the USA, England, and Australia, the literature provides robust results for heterogeneity within university graduates, at least for subpopulations. Graduating from a top or elite university leads to a wage premium. Chapter 4 addresses this question for Germany, a country with a rather flat university hierarchy compared to the countries named before. I sort universities into “better” and “worse” with the help of two rankings: (i) the worldwide QR ranking of top universities and (ii) a revealed preferences and acceptance (RPA) ranking based on high school GPAs. The QS ranking is publicly known, promoted by the institute, and the universities (if they scored high). The RPA ranking is not communicated, only has a mid-level correlation with the QR ranking, and is only based on individuals’ pre-ability measured by high school GPA. While both rankings reveal a wage premium, there are some interesting differences. The QS ranking shows a wage premium that is especially pronounced within the first year after graduation, whereas the RPA ranking yields significant effects five years after graduation. Because of the inherent differences in the rankings, the QS ranking could include more signaling than actual human capital accumulation compared to the RPA ranking. Since the RPA ranking requires up to five years to outweigh the QS ranking, students from high-ranked RPA universities need time to show their employers that they have higher skills and, thus, to earn more money. However, this is not testable with the data and should be looked at in more depth in the future. A further notable result is that women get a higher wage premium from top universities compared to men.

The second part of the dissertation relates to the e-learning literature and deals with the question of whether e-learning practice with knowledge of correct response can improve students’ exam grades and whether one can increase the students’ learning gains by giving additional hints in multiple-try feedback scenarios with knowledge of correct response. All three chapters in this part focus on applied settings, i.e., the observation of real university courses rather than of laboratory data. Chapter 5 combines rewarding and non-rewarding e-learning exercises. Students were allowed to participate in three midterm exams during the semester, which rewarded students with extra points in the exam if they performed well. After that, students could keep practicing with these midterms. Further, students could test themselves using an application ("a matrix a day") for the course's specific topic. Results suggest that participating and performing well in the midterms and exercises lead to better grades. Moreover, being good at the exercises in the application was also associated with better grades, while the number of submissions seems irrelevant. The results stay robust conditional on pre-ability measures, different scales of motivation and goals, as well as personality traits, suggesting causal validity.

To give additional evidence to the practice effect revealed in Chapter 5, Chapter 6 uses another setting with non-rewarded but weekly e-learning exercises to confirm the positive impact of e-learning exercises on exam scores. For this sample, I confirm the participation but not the performance effect. Here, however, is the correlation between the two variables much higher compared to the chapter before, and the number of individuals is lower. Thus, there might not have been enough data to capture both impacts accurately.

Since we found positive e-learning practice effects in Chapters 5 and 6, Chapter 7 then analyzes whether one can increase students' learning gains during an e-learning exercise. Therefore, we used a randomized within variation experiment given only knowledge of correct response to half of the students, while the others received additional hints on how to solve the exercises. The chapter finds that students receiving additional hints achieved more points during the learning phase and a week later in previous exam questions. The general practice effects on exam grades are shown in this chapter as well. Therefore, the dissertation highlights that students should get the possibility for additional practice with knowledge of correct response and additional hints. This should help to foster students' knowledge in university classes.

As a general summary, this Ph.D. thesis underscores the importance of educational qualification for economic outcomes and that the same education qualification affects women and men differently. Further, the second part emphasizes possibilities in e-learning environments to improve learning processes that enhance learning outcomes.

Bibliography

Acee, T. W. & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *Journal of Experimental Education*, *78*(4), 487–512. (Cited on page 107).

Acemoglu, D. & Angrist, J. D. (2000). How large are human-capital externalities? Evidence from compulsory schooling laws. *NBER Macroeconomics Annual*, *15*(2000), 9–59. (Cited on page 2).

Allensworth, E., Nomi, T., Montgomery, N. & Lee, V. E. (2009). College preparatory curriculum for all: Academic consequences of requiring Algebra and English I for ninth graders in chicago. *Educational Evaluation and Policy Analysis*, *31*(4), 367–391. (Cited on page 49).

Alpert, W. T., Couch, K. A., Harmon & R, O. (2016). A randomized assessment of online learning. *American Economic Review: Papers & Proceedings*, *106*(5), 378–382. (Cited on page 8).

Altonji, J. G. (1995). The effects of high school curriculum on education and labor market outcomes. *Journal of Human Resources*, *30*(3), 409–438. (Cited on page 49).

Altonji, J. G., Blom, E. & Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual Review of Economics*, *4*(1), 185–223. (Cited on page 47).

Andrews, R. J., Li, J. & Lovenheim, M. F. (2016). Quantile treatment effects of college quality on earnings. *Journal of Human Resources*, *51*(1), 200–238. (Cited on page 74).

Anelli, M. (2016). The returns to elite college education: A quasi-experimental analysis. *IZA Discussion Paper*, (10192). (Cited on pages 74, 76).

Angrist, J. D. & Evans, W. N. (1999). Schooling and labor market consequences of the 1970 state abortion reforms. In S. W. Polachek (Editor), *Research in labor economics* (Volume 18, Pages 75–113). (Cited on page 3).

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, *106*(4), 979–1014. (Cited on pages 3, 87).

Angrist, J. D. & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. *Princeton University Press*, (March). (Cited on page 84).

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, *121*(1-2), 343–375. (Cited on pages 19, 45, 47).

Ardac, D. & Sezen, A. H. (2002). Effectiveness of computer-based chemistry instruction in enhancing the learning of content and variable control under guided versus unguided conditions. *Journal of Science Education and Technology*, *11*(1), 39–48. (Cited on page 187).

Ashenfelter, O. & Krueger, A. B. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review*, *84*(5), 1157–1173. (Cited on page 3).

Ashenfelter, O. & Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *Quarterly Journal of Economics*, *113*(1), 253–284. (Cited on page 3).

Attali, Y. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers and Education*, *86*, 260–267. (Cited on pages 167, 169, 171, 186, 188, 189).

Aughinbaugh, A. (2012). The effects of high school math curriculum on college attendance: Evidence from the NLSY97. *Economics of Education Review*, *31*(6), 861–870. (Cited on page 48).

Avery, C. N., Glickman, M. E., Hoxby, C. M. & Metrick, A. (2013). A revealed preference ranking of U.S. colleges and universities. *Quarterly Journal of Economics*, *128*(1), 425–467. (Cited on page 74).

Azevedo, R. & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, *13*(2), 111–127. (Cited on pages 166, 188).

Bailey, T. H. & Phillips, L. J. (2016). The influence of motivation and adaptation on students' subjective well-being, meaning in life and academic performance. *Higher Education Research and Development*, *35*(2), 201–216. (Cited on pages 106, 141).

Baillet, F., Franken, A. & Weber, A. (2017). DZHW graduate panel 2009. Data and methods report on the graduate panel 2009 (1st and 2nd survey waves). *Data and Methods Report fdz.DZHW*, 1–34. (Cited on pages 20, 46, 59, 80).

Baillet, F., Franken, A. & Weber, A. (2019). DZHW-Absolventenpanel 2005. Daten- und Methodenbericht zu den Erhebungen der Absolvent(inn)enkohorte 2005 (1., 2. und 3. Befragungswelle). *Daten- und Methodenbericht fdz.DZHW*, 1–49. (Cited on pages 20, 46, 59, 80).

Baker, R., Evans, B., Li, Q. & Cung, B. (2019). Does inducing students to schedule lecture watching in online classes improve their academic performance? An experimental analysis of a time management intervention. *Research in Higher Education*, *60*(4), 521–552. (Cited on pages 107, 142).

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A. & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238. (Cited on page 166).

Barkley, E. F. (2010). *Student engagement techniques: A handbook for college faculty.* San Francisco, CA: Jossey-Bass. (Cited on page 9).

Barro, R. J. (2015). Human capital and growth. *American Economic Review*, *91*(2), 85–88. (Cited on page 5).

Becker, A., Deckers, T., Dohmen, T., Falk, A. & Kosse, F. (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics*, *4*(1), 453–478. (Cited on page 116).

Becker, S. O., Fernandes, A. & Weichselbaumer, D. (2019). Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment. *Labour Economics*, *59*, 139–152. (Cited on page 30).

Belloni, A., Chernozhukov, V. & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608–650. (Cited on pages 118, 123, 154, 156).

Belman, D. & Heywood, J. S. (1991). Sheepskin effects in the returns to education in a developing country. *Review of Economics and Statistics*, *73*(4), 720–724. (Cited on pages 5, 12, 93).

Berlingieri, F. & Zierahn, U. (2014). Field of study, qualification mismatch, and wages: Does sorting matter? *SSRN Electronic Journal*, *14-076*, 1–41. (Cited on page 45).

Bertrand, M., Duflo, E. & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249–275. (Cited on page 58).

Bettinger, E. P., Fox, L., Loeb, S. & Taylor, E. S. (2017). Virtual classrooms: How online college courses affect student success. *American Economic Review*, *107*(9), 2855–2875. (Cited on page 8).

Biewen, M. & Schwerter, J. (2019). Does more math in high school increase the share of female STEM workers? Evidence from a curriculum reform. *IZA Discussion Paper*, (12236), 1–33. (Cited on page 19).

Biewen, M. & Seifert, S. (2018). Potential parenthood and career progression of men and women – A simultaneous hazards approach. *B.E. Journal of Economic Analysis and Policy*, *18*(2), 1–22. (Cited on page 30).

Birch, E. R., Li, I. & Miller, P. W. (2009). The influences of institution attended and field of study on graduates' starting salaries. *Australian Economic Review*, *42*, 42–63. (Cited on pages 74, 75, 76, 79).

Black, D. A. & Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, *121*(1-2), 99–124. (Cited on page 74).

Black, D. A. & Smith, J. A. (2006). Estimating the returns to college quality with multiple proxies for quality. *Journal of Labor Economics*, *3*, 701–728. (Cited on page 74).

Bowen, W. G., Chingos, M. M., Lack, K. A. & Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management*, *33*(1), 94–111. (Cited on page 8).

Brand, J. E. & Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research*, *35*(3), 749–770. (Cited on pages 74, 76).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. (Cited on page 122).

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984). *Classification and regression trees*. CRC press. (Cited on page 121).

Brenoe, A. & Zölitz, U. (2020). Exposure to more female peers widens the gender gap in STEM participation. *Journal of Labor Economics*, 1–57. (Cited on pages 47, 50, 67, 69).

Brewer, D. J., Eide, E. R. & Ehrenberg, R. G. (1999). Does it pay to attend an elite private college? Cross-cohort evidence on the effects of college type on earnings college. *Journal of Human Resources*, *34*(1), 104–123. (Cited on pages 74, 76).

Broadbent, J. & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education*, *27*, 1–13. (Cited on pages 9, 104, 106, 140).

Broecke, S. (2013). Does offering more science at school increase the supply of scientists? The impact of offering triple science at GCSE on subsequent educational choices and outcomes. *Education Economics*, *21*(4), 325–342. (Cited on pages 45, 48).

Brown, B. W. & Liedholm, C. E. (2002). Can web courses replace the classroom in principles of microeconomics? *American Economic Review*, *92*(2), 444–448. (Cited on pages 8, 104, 140).

Brown, M. G. (2016). Blended instructional practice: A review of the empirical literature on instructors' adoption and use of online tools in face-to-face teaching. *Internet and Higher Education*, *31*, 1–10. (Cited on pages 7, 104, 106, 140).

Bundesagentur für Arbeit. (2019). Blickpunkt Arbeitsmarkt - MINT-Berufe. *Berichte: Blickpunkt Arbeitsmarkt*, (August), 1–39. (Cited on pages 15, 28, 60).

Burke, R. J. (2007). Women and minorities in STEM: A primer. In R. J. Burke & M. C. Mattis (Editors), *Women and minorities in science, technology, engineering and mathematics: Upping the numbers* (Chapter 1, Pages 3–27). (Cited on page 15).

Buser, T., Peter, N. & Wolter, S. C. (2017). Gender, competitiveness, and study choices in high school: Evidence from Switzerland. *American Economic Review*, *107*(5), 125–130. (Cited on pages 36, 50, 67).

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology*, *36*(5), 1118–1133. (Cited on pages 104, 140).

Butler, A. C., Karpicke, J. D. & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*(4), 918–928. (Cited on page 129).

Card, D. (1995). Earnings, schooling, and ability revisited. In S. W. Polachek (Editor), *Research in labor economics* (Volume 14, Pages 23–48). (Cited on pages 3, 87).

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Editors), *Handbook of Labor Economics* (Chapter 30, Volume 3, Pages 1801–1863). Elsevier. (Cited on pages 2, 3).

Card, D. & Payne, A. A. (2017). High school choices and the gender gap in STEM. *NBER Working Paper*, (23769), 1–24. (Cited on page 48).

Carrell, S. E., Page, M. E. & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144. (Cited on page 45).

Carroll, D. (2014). An investigation of the relationship between university rankings and graduate starting wages. *Journal of Institutional Research*, *19*(1), 46–54. (Cited on pages 74, 75, 76, 79).

Carroll, D., Heaton, C. & Tani, M. (2018). Does it pay to graduate from an 'elite' university in Australia? *IZA Discussion Paper*, (11477). (Cited on pages 74, 75, 76, 79).

Cech, E., Rubineau, B., Silbey, S. & Seron, C. (2011). Professional role confidence and gendered persistence in engineering. *American Sociological Review*, *76*(5), 641–666. (Cited on page 50).

Ceci, S. J. & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current Directions in Psychological Science*, *19*(5), 275–279. (Cited on pages 18, 45).

Cervone, D. (2012). MathJax: a platform for mathematics on the Web. *Notices of the AMS*, *59*(2), 312–316. (Cited on page 174).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Li, Y. (2020). *Xgboost: Extreme gradient boosting.* R package version 1.1.1.1. (Cited on page 123).

Chen, W., Grove, W. A. & Hussey, A. (2012). The payoff to school selectivity: An application of dale and krueger's method to MBA programs. *Economics Letters*, *116*(2), 247–249. (Cited on pages 74, 77).

Cheryan, S. (2012). Understanding the paradox in math-related fields: Why do some gender gaps remain while others do not? *Sex Roles*, *66*, 184–190. (Cited on pages 45, 50).

Clariana, R. B. & Koul, R. (2005). Multiple-try feedback and higher-order learning outcomes. *International Journal of Instructional Media*, *32*(3), 239–245. (Cited on pages 167, 188).

Clark, C. M. & Bjork, R. A. (2012). When and why introducing difficulties and errors can enhance instruction. In V. A. Benassi, C. E. Overson & C. M. Hakala (Editors), *Applying science of learning in education: Infusing psychological science into the curriculum* (Pages 20–30). (Cited on page 129).

Coates, D., Humphreys, B. R., Kane, J. & Vachris, M. A. (2004). "No significant distance" between face-to-face and online instruction: Evidence from principles of economics. *Economics of Education Review*, *23*(5), 533–546. (Cited on pages 8, 104, 140).

Coates, H. (2006). *Student engagement in campus based and online education: Univevrsity connections*. London: Routledge. (Cited on page 9).

Condron, D. J., Becker, J. H. & Bzhetaj, L. (2018). Sources of students' anxiety in a multidisciplinary social statistics course. *Teaching Sociology*, *46*(4), 346–355. (Cited on page 141).

Corbett, A. T. & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the sigchi conference on human factors in computing systems* (Pages 245–252). (Cited on page 186).

Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, *106*(6), 1691–1730. (Cited on page 50).

Cortes, K. E., Goodman, J. & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose-algebra. *Journal of Human Resources*, *50*(1), 108–158. (Cited on page 49).

Dale, S. B. & Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *Quarterly Journal of Economics*, *117*(4), 1491–1527. (Cited on pages 74, 77).

Dale, S. B. & Krueger, A. B. (2014). Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of Human Resources*, *49*(2), 323–358. (Cited on pages 74, 77).

Danbold, F. & Huo, Y. J. (2017). Men's defense of their prototypicality undermines the success of women in STEM initiatives. *Journal of Experimental Social Psychology*, *72*, 57–66. (Cited on pages 15, 36, 37, 45).

Darolia, R., Koedel, C., Main, J. B., Ndashimye, J. F. & Yan, J. (2019). High school course access and postsecondary STEM enrollment and attainment. *Educational Evaluation and Policy Analysis*, *42*(1), 22–45. (Cited on page 49).

De Philippis, M. (2017). STEM graduates and secondary school curriculum: Does early exposure to science matter? *Bank of Italy Temi di Discussione (Working Paper)*, (1107). (Cited on pages 47, 49).

de Walque, D. (2010). Education, information, and smoking decisions: Evidence from smoking histories in the United States, 1940-2000. *Journal of Human Resources*, *45*(3), 682–717. (Cited on page 4).

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics*, *88*(9-10), 1697–1720. (Cited on pages 4, 87).

Del Carpio, L. & Guadalupe, M. (2018). More women in tech? Evidence from a field experiment addressing social identity. *IZA Discussion Paper*, (11876), 1–49. (Cited on page 50).

Dickson, M. & Harmon, C. (2011). Economic returns to education: What we know, what we don't know, and where we are going – some brief pointers. *Economics of Education Review*, *30*(6), 1118–1122. (Cited on page 2).

Domina, T. & Saldana, J. (2012). Does raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982-2004. *American Educational Research Journal*, *49*(4), 685–708. (Cited on pages 46, 48).

Ehrenberg, R. G. (2010). Analyzing the factors that influence persistence rates in STEM field, majors: Introduction to the symposium. *Economics of Education Review*, *29*(6), 888–891. (Cited on page 45).

Elliot, A. J. & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, *100*(3), 613–628. (Cited on pages 116, 150).

Enman, M. & Lupart, J. (2000). Talented female students' resistance to science: An exploratory study of post-secondary achievement motivation, persistence, and epistemological characteristics. *High Ability Studies*, *11*(2), 161–178. (Cited on page 19).

Fehse, S. & Kerst, C. (2007). Arbeiten unter Wert? Vertikal und horizontal inadäquate Beschäftigung von Hochschulabsolventen der Abschlussjahrgänge 1997 und 2001. *Beiträge zur Hochschulforschung*, *29*(1), 72–98. (Cited on page 24).

Figlio, D., Rush, M. & Yin, L. (2013). Is it live or is it internet? Experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, *31*(4), 763–784. (Cited on pages 8, 104, 140).

Finney, S. J. & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161–186. (Cited on page 107).

Fischer, C., Baker, R., Li, Q., Orona, G. & Warschauer, M. (2019). Does online course-taking increase distal student success? Examining impacts on college graduation rates and time to degree. *AERA Online Paper Repository*, 1–16. (Cited on page 8).

Fischer, C., Zhou, N., Rodriguez, F., Warschauer, M. & King, S. (2019). Improving college student success in organic chemistry: Impact of an online preparatory course. *Journal of Chemical Education*, *96*(5), 857–864. (Cited on pages 9, 104, 108, 140, 142, 188).

Franceschini, G., Galli, S., Chiesi, F. & Primi, C. (2014). Implicit gender-math stereotype and women's susceptibility to stereotype threat and stereotype lift. *Learning and Individual Differences*, *32*, 273–277. (Cited on pages 45, 50, 67).

Frederick, S., Loewenstein, G. & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401. (Cited on pages 116, 150).

Friedman-Sokuler, N. & Justman, M. (2016). Gender streaming and prior achievement in high school science and mathematics. *Economics of Education Review*, *53*, 230–253. (Cited on pages 18, 36, 45).

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer. (Cited on pages 121, 154).

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U. & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, *48*, 67–84. (Cited on pages 116, 150).

Ge, S., Isaac, E. & Miller, A. (2018). Elite schools and opting-in: Effects of college selectivity on career and family outcomes. *NBER Working Paper*, (25315), 1–42. (Cited on page 77).

Gneezy, U., Niederle, M. & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, *118*(3), 1049–1074. (Cited on pages 50, 67).

Goldin, C. & Katz, L. F. (2008). *The race between education and technology.* Cambridge MA.: Harvard University Press. (Cited on page 5).

Goodman, J. (2019). The labor of division: Returns to compulsory high school math coursework. *Journal of Labor Economics*, *37*(4), 1141–1182. (Cited on page 50).

Görlitz, K. & Gravert, C. (2016). The effects of increasing the standards of the high school curriculum on school dropout. *Applied Economics*, *48*(54), 5314–5328. (Cited on pages 51, 67).

Görlitz, K. & Gravert, C. (2018). The effects of a high school curriculum reform on university enrollment and the choice of college major. *Education Economics*, *26*(3), 321–336. (Cited on page 51).

Greenwood, J., Guner, N., Kocharkov, G. & Santos, C. (2014). Marry your like: Assortative mating and income inequality. *American Economic Review*, *104*(5), 348–353. (Cited on page 5).

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911–922. (Cited on page 45).

Grossman, M. (2006). Education and nonmarket outcomes. In E. Hanushek & F. Welch (Editors), *Handbook of the Economics of Education* (Chapter 10, Volume 1, Pages 577–633). Elsevier. (Cited on pages 2, 3, 4, 6).

Gylfason, T. (2001). Natural resources, education, and economic development. *European Economic Review*, *45*(4-6), 847–859. (Cited on page 5).

Hartog, J., Sun, Y. & Ding, X. (2010). University rank and bachelor's labour market positions in China. *Economics of Education Review*, *29*(6), 971–979. (Cited on page 75).

Hattie, J., Gan, M. & Brooks, C. (2016). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Editors), *Handbook of research on learning and instruction* (Chapter 14). (Cited on page 166).

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. (Cited on page 166).

Heilbronner, N. N. (2013). The STEM pathway for women: What has changed? *Gifted Child Quarterly*, *57*(1), 39–55. (Cited on pages 19, 36).

Helliwell, J. F. & Putnam, R. D. (2007). Education and social capital. *Eastern Economic Journal*, *33*(1), 1–19. (Cited on page 5).

Henderson, D. J., Polachek, S. W. & Wang, L. (2011). Heterogeneity in schooling rates of return. *Economics of Education Review*, *30*(6), 1202–1214. (Cited on page 3).

Hjalmarsson, R., Holmlund, H. & Lindquist, M. J. (2015). The effect of education on criminal convictions and incarceration: Causal evidence from micro-data. *Economic Journal*, *125*(587), 1290–1326. (Cited on page 4).

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. (Cited on page 122).

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *Review of Economics and Statistics*, *91*(4), 717–724. (Cited on pages 74, 76).

Honicke, T. & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review*, *17*, 63–84. (Cited on pages 106, 141).

Hübner, N., Wille, E., Cambria, J., Oschatz, K., Nagengast, B. & Trautwein, U. (2017). Maximizing gender equality by minimizing course choice options? Effects of obligatory coursework in math on gender differences in STEM. *Journal of Educational Psychology*, *109*(7), 993–1009. (Cited on pages 19, 47, 50, 51, 67, 69).

Hulleman, C. S., Schrager, S. M., Bodmann, S. M. & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*(3), 422–449. (Cited on page 116).

Hunt, J. (2016). Why do women leave science and engineering? *ILR Review*, *69*(1), 199–226. (Cited on pages 16, 18, 21, 24, 28, 37).

Hussain, I., McNally, S. & Telhaj, S. (2009). University quality and graduate wages in the UK. *IZA Discussion Paper*, (4043), 1–26. (Cited on pages 74, 76).

Ivanova, M. & Stein, P. (2013). Bachelorabschluss: Endstation von Chemikerinnen? Einstellungen von Studierenden der Chemie zum Studium und zur beruflichen Karriere - Ergebnisse einer Befragung an zwölf ausgewählten Hochschulen in Deutschland. In U. Pascher & P. Stein (Editors), *Akademische Karrieren von Naturwissenschaftlerinnen gestern und heute* (1st edition, Pages 125–149). Springer Fachmedien. (Cited on page 18).

Jacob, B., Dynarski, S., Frank, K. & Schneider, B. (2017). Are expectations alone enough? Estimating the effect of a mandatory college-prep curriculum in Michigan. *Educational Evaluation and Policy Analysis*, *39*(2), 333–360. (Cited on page 49).

Jaehnig, W. & Miller, M. L. (2007). Feedback types in programmed instruction: A systematic review. *Psychological Record*, *57*(2), 219–232. (Cited on page 166).

Jaggars, S. S. (2014). Choosing between online and face-to-face courses: Community college student voices. *American Journal of Distance Education*, *28*(1), 27–38. (Cited on page 9).

Jaggars, S. S. & Xu, D. (2016). How do online course design features influence student performance? *Computers and Education*, *95*, 270–284. (Cited on page 9).

Jansen, K. & Pascher, U. (2013). “Und dann hat man keine Zeit mehr für Familie oder so.” - Wissenschaftsorientierung und Zukunftsvorstellungen von Bachelorstudentinnen chemischer Studiengänge. In U. Pascher & P. Stein (Editors), *Akademische Karrieren von Naturwissenschaftlerinnen gestern und heute* (1st edition, Pages 151–192). Springer Fachmedien. (Cited on pages 18, 37).

Jansen, M., Scherer, R. & Schroeders, U. (2015). Students’ students’ self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, *41*, 13–24. (Cited on page 48).

Jia, N. (2016). Do stricter high school math requirements raise college STEM attainment? *Working paper, Mimeo*, 1–39. (Cited on pages 47, 49).

Joensen, J. S. & Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, *44*(1), 171–198. (Cited on page 50).

Joensen, J. S. & Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *Economic Journal*, *126*(593), 1129–1163. (Cited on pages 47, 50).

Jones, L. E., Schoonbroodt, A. & Tertilt, M. (2010). Fertility theories: Can they explain the negative fertility-income relationship? In *Demography and the Economy* (Pages 43–100). NBER Chapters. National Bureau of Economic Research, Inc. (Cited on page 5).

Joy, L. (2000). Do colleges shortchange women? Gender differences in the transition from college to work. *American Economic Review*, *90*(2), 471–475. (Cited on page 17).

Joyce, T., Crockett, S., Jaeger, D. A., Altindag, O. & O'Connell, S. D. (2015). Does classroom time matter? *Economics of Education Review*, *46*, 64–77. (Cited on page 8).

Jung, J. & Lee, S. J. (2016). Influence of university prestige on graduate wage and job satisfaction: The case of South Korea. *Journal of Higher Education Policy and Management*, *38*(3), 297–315. (Cited on pages 74, 76).

Justman, M. & Mendez, S. J. (2018). Gendered choices of STEM subjects for matriculation are not driven by prior differences in mathematical achievement. *Economics of Education Review*, *64*(3), 282–297. (Cited on pages 15, 45, 46, 48).

Kahn, S. & Ginther, D. K. (2015). Are recent cohorts of women with engineering bachelors less likely to stay in engineering? *Frontiers in Psychology*, *6*(1144), 1–15. (Cited on pages 16, 18).

Kahn, S. & Ginther, D. K. (2018). Women and science, technology, engineering, and mathematics (STEM): Are differences in education and careers due to stereotypes, interests, or family? In S. L. Averett, L. M. Argys & S. D. Hoffmann (Editors), *The oxford handbook of women and the economy* (Chapter 31). Oxford: Oxford University Press. (Cited on page 47).

Kemptner, D., Jürges, H. & Reinhold, S. (2011). Changes in compulsory schooling and the causal effect of education on health: Evidence from Germany. *Journal of Health Economics*, *30*(2), 340–354. (Cited on pages 4, 87).

Kirschner, P. A., Martens, R. L. & Strijbos, J.-W. (2004). CSCL in higher education? A framework for designing multiple collaborative environments. In J.-W. Strijbos, P. A. Kirschner & R. L. Martens (Editors), *What we know about CSCL* (Chapter 1, Pages 3–30). Springer, Dordrecht. (Cited on page 187).

Kirschner, P. A., Sweller, J. & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. (Cited on page 186).

Kizilcec, R. F., Pérez-Sanagustín, M. & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers and Education*, *104*, 18–33. (Cited on pages 104, 140).

Kleinberg, E. M. (1996). An overtraining-resistant stochastic modeling method for pattern recognition. *Annals of Statistics*, *24*(6), 2319–2349. (Cited on page 122).

Kling, J. R. (2001). Interpreting instrumental variables estimates of the returns to schooling. *Journal of Business and Economic Statistics*, *19*(3), 358–364. (Cited on page 87).

Kluger, A. N. & DeNisi, A. S. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. (Cited on page 166).

Koenker, R. & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50. (Cited on pages 123, 124).

Kokkelenberg, E. C. & Sinha, E. (2010). Who succeeds in STEM studies? An analysis of Binghamton university undergraduate students. *Economics of Education Review*, *29*(6), 935–946. (Cited on page 45).

Komarraju, M. & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences*, *25*, 67–72. (Cited on pages 106, 141).

Kuhn, M. (2020). Caret: Classification and regression training. (Cited on page 121).

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2), 211–232. (Cited on pages 166, 187).

Kulik, J. A. & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97. (Cited on page 166).

Kursa, M. B. & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. (Cited on pages 122, 154).

Lafortune, J. (2013). Making yourself attractive: Pre-marital investments and the returns to education in the marriage market. *American Economic Journal: Applied Economics*, *5*(2), 151–178. (Cited on page 5).

Lane, A. M., Hall, R. & Lane, J. (2004). Self-efficacy and statistics performance among sport studies students. *Teaching in Higher Education*, *9*(4), 435–448. (Cited on page 107).

Lemke, R. J. & Rischall, I. C. (2003). Skill, parental income, and IV estimation of the returns to schooling. *Applied Economics Letters*, *10*(5), 281–286. (Cited on page 87).

Lepper, M. R. & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Editor), *Improving academic achievement: Impact of psychological factors on education* (Chapter 7, Pages 135–158). Cambridge, MA: Academic Press. (Cited on page 186).

Levine, P. B. & Zimmerman, D. J. (1995). The benefit of additional high-school math and science classes for young men and women. *Journal of Business & Economic Statistics*, *13*(2), 137–149. (Cited on page 47).

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. (Cited on pages 122, 154).

Lindahl, L. & Regnér, H. (2005). College choice and subsequent earnings: Results using Swedish sibling data. *Scandinavian Journal of Economics*, *107*(3), 437–457. (Cited on page 74).

Lochner, L. (2004). Education, work, and crime: A human capital approach. *International Economic Review*, *45*(3), 811–843. (Cited on page 4).

Lochner, L. & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, *94*(1), 155–189. (Cited on page 4).

Loewenstein, G., O'Donoghue, T. & Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, *118*(4), 1209–1248. (Cited on page 116).

Long, M. C. (2008). College quality and early adult outcomes. *Economics of Education Review*, *27*(5), 588–602. (Cited on pages 74, 76).

Lubinski, D. & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*(4), 316–345. (Cited on page 19).

Macher, D., Papousek, I., Ruggeri, K. & Paechter, M. (2015). Statistics anxiety and performance: Blessings in disguise. *Frontiers in Psychology*, *6*(1116), 4–7. (Cited on page 107).

Machin, S., Salvanes, K. G. & Pelkonen, P. (2012). Education and mobility. *Journal of the European Economic Association*, *10*(2), 417–450. (Cited on page 4).

Mackinnon, J. G. & Webb, M. D. (2017). Pitfalls when estimating treatment effects using clustered data. *Political Methodologist*, *24*(2), 20–31. (Cited on page 58).

Malamud, O. (2011). Discovering one's talent: Learning from academic specialization. *ILR Review*, *64*(2), 375–405. (Cited on pages 3, 6, 11, 194).

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U. & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471–491. (Cited on page 116).

Marsh, H. W. & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, *81*(1), 59–77. (Cited on page 116).

McKenzie, K. & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development*, *20*(1), 21–33. (Cited on pages 106, 141).

Meier, S. & Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, *2*(1), 193–210. (Cited on page 116).

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics*, *14*(4), 415–429. (Cited on page 187).

Milla, J. (2017). The context-bound university selectivity premium. *IZA Discussion Paper*, (11025). (Cited on page 74).

Milligan, K., Moretti, E. & Oreopoulos, P. (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. *Journal of Public Economics*, *88*(9-10), 1667–1695. (Cited on pages 4, 87).

Mincer, J. A. (1974). The human capital earnings function. In J. A. Mincer (Editor), *Schooling, experience, and earnings* (Pages 83–96). NBER. (Cited on page 2).

Monks, J. (2000). The returns to individual and college characteristics: Evidence from the national longitudinal survey of youth. *Economics of Education Review*, *19*(3), 279–289. (Cited on pages 74, 77).

Morgan, L. A. (2000). Is engineering hostile to women? An analysis of data from the 1993 national survey of college graduates. *American Sociological Review*, *65*(2), 316–321. (Cited on pages 16, 18).

Morgan, S. L., Gelbgiser, D. & Weeden, K. A. (2013). Feeding the pipeline: Gender, occupational plans, and college major selection. *Social Science Research*, *42*(4), 989–1005. (Cited on page 50).

Narciss, S. (2006). Modelle zu den Bedingungen und Wirkungen von Feedback in Lehr-Lernsituationen. In H. Ditton & A. Müller (Editors), *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Chapter 2.3, Pages 43–83). Münster: Waxmann. (Cited on page 166).

O'Flaherty, J. & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *Internet and Higher Education*, *25*, 85–95. (Cited on pages 8, 9, 104, 140).

OECD. (2007). Women in science, engineering and technology (SET): Strategies for a global workforce. *Workshop Summary.* (Cited on page 45).

OECD. (2010). OECD information technology outlook 2010, 1–299. (Cited on page 45).

OECD. (2016). OECD factbook 2015-2016: Economic, environmental and social statistics, 1–223. (Cited on page 6).

OECD. (2017). Education at a glance 2017: OECD indicators, 1–452. (Cited on pages 2, 24).

Oechsle, M., Knauf, H., Maschetzke, C. & Rosowski, E. (2009). Wie tragfähig ist die Studien- und Berufswahl? Biographische Verläufe und Orientierungsprozesse nach dem Abitur. In M. Oechsle, H. Knauf, C. Maschetzke & E. Rosowski (Editors), *Abitur und was dann?* (Chapter 8, Pages 283–324). (Cited on page 17).

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, *91*(11-12), 2213–2229. (Cited on pages 3, 4, 87).

Oreopoulos, P. & Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, *25*(1), 159–184. (Cited on pages 3, 4, 5).

Paechter, M., Maier, B. & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers and Education*, *54*(1), 222–229. (Cited on pages 104, 106, 140, 141).

Park, H., Behrman, J. R. & Choi, J. (2018). Do single-sex schools enhance students' STEM (science, technology, engineering, and mathematics) outcomes? *Economics of Education Review*, *62*, 35–47. (Cited on page 107).

Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P. & Warschauer, M. (2018). Understanding student procrastination via mixture models. *Proceedings of the 11th International Conference on Educational Data Mining (EDM)*, 187–197. (Cited on page 142).

Perez, T., Cromley, J. G. & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315–329. (Cited on page 48).

Perna, L. W. (2004). Understanding the decision to enroll in graduate school: Sex and racial/ethnic group differences. *Journal of Higher Education*, *75*(5), 487–527. (Cited on page 18).

Polachek, S. W. (1981). Occupational self-selection: A human capital approach to sex differences in occupational structure. *Review of Economics and Statistics*, *63*(1), 60–69. (Cited on pages 18, 37).

Powdthavee, N., Lekfuangfu, W. N. & Wooden, M. (2015). What's the good of education on our overall quality of life? A simultaneous equation model of education and life satisfaction for Australia. *Journal of Behavioral and Experimental Economics*, *54*, 10–21. (Cited on page 4).

Preston, A. E. (1994). Why have all the women gone? A study of exit of women from the science and engineering professions. *American Economic Review*, *84*(5), 1446–1462. (Cited on pages 16, 18).

Preston, A. E. (2004). *Leaving science: occupational exit from scientific careers*. New York: Russel Sage Foundation. (Cited on page 18).

Pritchett, L. (2006). Does learning to add up add up? The returns to schooling in aggregate data. In E. Hanushek & F. Welch (Editors), *Handbook of the Economics of Education* (Chapter 11, Volume 1, Pages 635–695). Elsevier. (Cited on page 5).

Proske, A., Körndle, H. & Narciss, S. (2012). Interactive learning tasks. In N. M. Seel (Editor), *Encyclopedia of the sciences of learning* (Pages 1606–1610). Boston, MA: Springer US. (Cited on page 166).

Qualtrics. (2020). Qualtrics®. Provo, Utah, USA: Qualtrics. (Cited on page 174).

Rimfeld, K., Kovas, Y., Dale, P. S. & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology*, *111*(5), 780–789. (Cited on pages 106, 141).

Robst, J. (2007). Education and job match: The relatedness of college major and work. *Economics of Education Review*, *26*(4), 397–407. (Cited on page 24).

Rodriguez, F., Fischer, C., Zhou, N., Warschauer, M. & Massimelli, J. (2016). Spacing and self-testing strategies are positively associated with learning in an upper-division microbiology course. *SN Social Sciences*, *32*, 1–5. (Cited on pages 107, 142).

Rodriguez, F., Kataoka, S., Rivas, J. M., Kadandale, P., Nili, A. & Warschauer, M. (2018). Do spacing and self-testing predict learning outcomes? *Active Learning in Higher Education*. (Cited on pages 107, 142).

Rodriguez, F., Rivas, M. J., Matsumura, L. H., Warschauer, M. & Sato, B. K. (2018). How do students study in STEM courses? Findings from a light-touch intervention and its relevance for underrepresented students. *PLOS ONE*, *13*(7), 1–20. (Cited on page 107).

Roediger III, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. (Cited on page 129).

Roodman, D., MacKinnon, J. G., Nielsen, M. Ø. & Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *Stata Journal*, *19*(1), 4–60. (Cited on pages 63, 65, 66, 71, 72).

Rose, H. & Betts, J. R. (2004). The effect of high school courses on earnings. *Review of Economics and Statistics*, *86*(2), 497–513. (Cited on page 50).

Sassler, S., Glass, J., Levitte, Y. & Michelmore, K. M. (2017). The missing women in STEM? Assessing gender differentials in the factors associated with transition to first jobs. *Social Science Research*, *63*, 192–208. (Cited on pages 16, 45).

Schavan, A. (1999). Stellungnahme des Ministeriums für Kultur, Jugend und Sport, 1–7. (Cited on page 45).

Schupp, J. & Gerlitz, J.-Y. (2014). Big five inventory – SOEP (BFI-S). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. (Cited on pages 116, 150).

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. (Cited on page 166).

Simard, C., Henderson, A. D., Gilmartin, S. K., Schiebinger, L. & Whitney, T. (2013). *Climbing the technical ladder: Obstacles and solutions for mid-level women in technology*. (Cited on pages 15, 37).

Sloboda, J. A., Davidson, J. W., Howe, M. J. A. & Moore, D. G. (1996). The role of practice in the development of performing musicians. *British Journal of Psychology*, *87*(2), 287–309. (Cited on page 128).

Spitz-Oener, A. & Priesack, K. (2018). STEM occupations and the evolution of the German wage structure. *Working paper, Mimeo*, 1–64. (Cited on page 47).

Stephens, M. & Yang, D. Y. (2014). Compulsory education and the benefits of schooling. *American Economic Review*, *104*(6), 1777–1792. (Cited on page 87).

Sweller, J. (1999). *Instructional design in technical areas.* Camberwell, Vic: ACER Press. (Cited on page 187).

Sweller, J., van Merrienboer, J. J. G. & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. (Cited on page 187).

Thai, N. T. T., De Wever, B. & Valcke, M. (2017). The impact of a flipped classroom design on learning performance in higher education: Looking for the best "blend" of lectures and guiding questions with feedback. *Computers and Education*, *107*, 113–126. (Cited on pages 8, 104, 106, 140).

Thomas, S. L. (2003). Longer-term economic effects of college selectivity and control. *Research in Higher Education*, *44*(3), 263–299. (Cited on page 74).

Thomas, S. L. & Zhang, L. (2005). Post-baccalaureate wage growth within four years of graduation: The effects of college quality and college major. *Research in Higher Education*, *46*(4), 437–459. (Cited on pages 74, 77).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. (Cited on page 121).

UNESCO. (2017). Literacy rates continue to rise from one generation to the next. *Unesco Institute for Statistics*, *45*, 1–13. (Cited on page 2).

Urminsky, O., Hansen, C. & Chernozhukov, V. (2016). Using double-lasso regression for principled variable selection. *Randomized Social Experiments eJournal.* (Cited on page 154).

Van der Kleij, F. M., Feskens, R. C. & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. (Cited on page 166).

Verhaest, D., Sellami, S. & van der Velden, R. (2017). Differences in horizontal and vertical mismatches across countries and fields of study. *International Labour Review*, *156*(1), 1–23. (Cited on page 24).

Walker, I. & Zhu, Y. (2011). Differences by degree: Evidence of the net financial rates of return to undergraduate study for England and Wales. *Economics of Education Review*, *30*(6), 1177–1186. (Cited on pages 5, 12, 93).

Walker, I. & Zhu, Y. (2017). University selectivity and the graduate wage premium: Evidence from the UK. *IZA Discussion Paper*, (10536). (Cited on pages 74, 77).

Wang, M.-T., Eccles, J. S. & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, *24*(5), 770–775. (Cited on page 18).

Watt, H. M. G. & Eccles, J. S. (2008). *Gender and occupational outcomes: Longitudinal assessment of individual, social, and cultural influences* (H. M. G. Watt & J. S. Eccles, Editors). (Cited on page 48).

Weinstein, R. (2017). University selectivity, initial job quality, and longer-run salary. *IZA Discussion Paper*, (10911), 1–38. (Cited on pages 74, 77, 88).

Wigfield, A. & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. (Cited on page 116).

Winkelmann, L. & Winkelmann, R. (1998). Why are the unemployed so unhappy? Evidence from panel data. *Economica*, *65*(257), 1–15. (Cited on page 4).

Wiswall, M. & Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies*, *82*(2), 791–824. (Cited on page 19).

World Bank. (2011). *Learning for all: Investing in people's knowledge and skills to promote development - world bank group education strategy 2020*. Washington, D.C.: World Bank Group. (Cited on page 2).

Xu, D. & Jaggars, S. S. (2014). Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. *Journal of Higher Education*, *85*(5), 633–659. (Cited on pages 8, 104, 140).

Xu, Y. J. (2013). Career outcomes of STEM and non-STEM college graduates: Persistence in majored-field and influential factors in career choices. *Research in Higher Education*, *54*(3), 349–382. (Cited on pages 8, 24).

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resouces*, *48*(3), 545–595. (Cited on page 19).

Zhong, H. (2015). Does a college education cause better health and health behaviours? *Applied Economics*, *47*(7), 639–653. (Cited on page 4).

Zhou, L., Lin, H. & Lin, Y.-C. (2016). Education, intelligence, and well-being: Evidence from a semiparametric latent variable transformation model for multiple outcomes of mixed types. *Social Indicators Research*, *125*(3), 1011–1033. (Cited on page 4).