Kinne, Lavinia

Research Report

Culture, feedback, and gender in education

ifo Beiträge zur Wirtschaftsforschung, No. 101

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Kinne, Lavinia (2023) : Culture, feedback, and gender in education, ifo Beiträge zur Wirtschaftsforschung, No. 101, ISBN 978-3-95942-124-9, ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München, München

This Version is available at:
https://hdl.handle.net/10419/274522

ifo
BEITRÄGE
zur Wirtschaftsforschung

101
2023

Culture, Feedback, and Gender in Education

Lavinia Kinne

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

ifo
BEITRÄGE
zur Wirtschaftsforschung

101
2023

Culture, Feedback, and Gender in Education

Lavinia Kinne

Herausgeber der Reihe: Clemens Fuest
Schriftleitung: Chang Woon Nam

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über
https://dnb.d-nb.de
abrufbar.

ISBN Nr. 978-3-95942-124-9

Druck: Kreiter Druck, Wolfratshausen

ifo Institut im Internet:
https://www.ifo.de

Culture, Feedback, and Gender in Education

Inaugural-Dissertation
Zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)

eingereicht an der
Ludwig-Maximilians-Universität München
2023

vorgelegt von

Lavinia Marie-Louise Kinne

Referent: Prof. Dr. Ludger Wößmann
Korreferentin: Prof. Dr. Larissa Zierow
Promotionsabschlussberatung: 12. Juli 2023

Datum der mündlichen Prüfung: 07.07.2023
Namen der Berichterstatter:innen: Prof. Dr. Ludger Wößmann
Prof. Dr. Larissa Zierow
Prof. Dr. Fabian Kosse

Preface

Lavinia Kinne prepared this study while she was working at the Center for the Economics of Education at the ifo Institute. The study was completed in March 2023 and accepted as doctoral thesis by the Department of Economics at LMU Munich. It consists of four distinct empirical essays that address various aspects of the economics of education. Chapters 2 and 3 show that patience and risk-taking as intertemporal preferences are closely related to differences in student achievement across and within countries. Chapter 3 hereby employs novel machine-learning methods to derive intertemporal preference measures from Facebook data. Using a field experiment among university students, chapter 4 shows that the ordering in which positive and negative performance feedback is provided, matters for study motivation and to some extent also exam performance. Finally, chapter 5 presents new insights to gender gaps in adult cognitive skills, showing that they are highly related to wages, especially at the top of the wage distribution.

Keywords: Patience, Risk-taking, Preferences, Culture, Intertemporal Decision-making, International Student Achievement, PISA, Student Achievement, Regions, Social Media, Facebook, Education, Feedback, Motivation, Performance, Gender Wage Gap, Skills, Numeracy, PIAAC

JEL-No: C93, D83, D91, I20, I21, I23, I24, J16, J24, Z10

"Solange ich schreibe, spreche ich zwar nicht,
aber ich schweige auch nicht."
(Kim de l'Horizon, 'Blutbuch', page 32.)

Acknowledgements

I am extremely grateful to my supervisor Ludger Wößmann who has guided me through these past years of preparing my dissertation. He managed to find the balance between close supervision and helping me whenever it was needed and letting me explore my own research interests and methods. I am deeply impressed by how he has been able to do this over the last years next to his excellent academic work and involvement in many important policy debates. I have highly benefited from being in his research group and especially from having him as a co-author on numerous projects.

A second very important advisor during my PhD was Larissa Zierow. She went from being a colleague to becoming my Postdoc supervisor, mentor, and friend. Besides being an excellent researcher, she is a very kind and thoughtful person, and I could always knock on her door when I had any type of problem. Also, she taught me and many others some of the most valuable skills in academia: how to engage, network, and have fun at conferences, and how to maintain good relationships in the profession. I also want to thank Fabian Kosse as my third supervisor, especially for also writing me a reference letter for the job market.

I was lucky to work with a series of inspiring co-authors during my PhD. Eric A. Hanushek and Philipp Lergetporer were part of my first PhD project and I learned a lot from them about how to do research. Rick also created such a nice atmosphere during our interactions and I am looking forward to working together more in the future. Pietro Sancassani joined on a follow-up project of this paper. I am very impressed by his thoroughness and calm when working on research (and in life). Michele Battisti and Alexandra Fedorets were my first co-authors outside of the Center which was such an enriching experience. They also became mentors over time and I am very grateful for their support.

I thank everyone at the ifo Center for the Economics of Education for the productive environment and insightful discussions. Ulrike Baldi-Cohrs and Franziska Binder were the most amazing team assistants one could imagine, and Benjamin 'Benji' Arold, Anna Katharina Wurm, Katharina Wedel, and Vera Freundl have become true friends over the years. I would also like to thank Sarah Gust, Andreas Leibing, and Jonas Jessen for the most wonderful two-conference week I will probably ever have.

Finally, I want to thank my family and friends for supporting me on this academic path. My largely academic German family sparked my interest in research early on and I never had to ask my parents twice to help me move anywhere in Europe for the next chapter of my life. The Italian family has always been an escape from this world and I want to especially thank my cousin Edoardo for our numerous talks about life, family, and research. Siegrid and Joachim, Paula, Benny, and Toni have been great friends throughout this journey.

Lavinia Kinne
March 2023

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Economics of Education

The relevance of education in the discipline of economics is rooted in its importance for many aggregate and individual economic outcomes. On the most aggregate level of economic outputs, education is highly related to economic growth (see e.g. Hanushek and Woessmann (2008, 2015)). So-called augmented neoclassical growth models highlight the importance of human capital as a direct outcome of education for increasing individuals' productivity on the labor market (Mankiw et al., 1992). Hence, *changes* in education should increase aggregate economic output which is assumed to be determined by capital and labor inputs (Solow, 1956). Instead, endogenous growth models emphasize the role of the *level* of education for fostering innovation in an economy in the long-run, e.g. through the continuous generation of new technologies even if no changes in education are observed (Lucas Jr, 1988; Romer, 1990; Howitt and Aghion, 1998). This concept then relates back to fundamental ideas from Schumpeter (1912) hypothesizing that innovation is a key driver of economic growth.

On the individual level, the skills acquired through education are known to be associated with better outcomes in a series of labor-market indicators (see e.g. Acemoglu and Autor (2011) and Hanushek et al. (2015)). The term *human capital* describes the stock of skills and other characteristics of a worker that determine her productivity. Investments in education increase human capital which is assumed to make more educated workers more productive (Becker, 1962). In the commonly used economic framework where the productivity of workers determines their wages, this should lead to higher wages for workers with higher skill levels.

The decision to invest in education is assumed to follow a series of individual considerations. On the one hand, there is an important trade-off between the benefits of education, both individual and societal, and the costs associated with education. These costs can be time or opportunity costs, e.g. in foregone earnings, as well as monetary costs of schooling materials or tuition fees (Mincer, 1958; Schultz, 1961; Becker, 1962). On the other hand, there is an inter-temporal trade-off since the costs occur immediately while the returns, e.g. on the labor market, are materialized much later in life (Mincer, 1958; Becker, 1964; Mincer, 1974). These conceptual considerations suggest that certain personal and/or cultural traits of individuals and societies can have an important impact on decisions to invest in the accumulation of further human capital.

How successful investments into education are in generating human capital then depends on the production process of the underlying skills. The existing literature about these so-called 'education production functions' has long focused on parental, school, and institutional

resources as input factors (Hanushek, 1986; Woessmann, 2016b). While these have shown to matter greatly for student outcomes, the deeper roots of differences in students' skills remained an open question. More recently, insights from behavioral economics have been added to the study of the determinants of skill production (Sutter et al., 2013; Golsteyn et al., 2014; Castillo et al., 2019; Figlio et al., 2019; Castillo et al., 2020). Inter-temporal personal and cultural traits such as patience and risk-taking might have an impact on human capital accumulation not only on the individual input level, but also by shaping institutions in a country an individual is schooled in. The latter is especially true since many input factors of the education production function are determined on more aggregated levels, i.e. of schools, regions, or even countries.

1.2 Cultural Determinants of Education Outcomes

The recent focus on behavioral and cultural dimensions to skill production has particularly studied patience and risk-taking as important determinants of educational outcomes. A higher level of patience can be interpreted as a lower rate of time discounting, implying that outcomes in the future are valued relatively more than current outcomes. This has important implications for the human capital investment decision where current costs are weighed against future benefits. Theoretically, this suggests that a higher level of patience is associated with higher investments into human capital. Empirically, Sutter et al. (2013), Golsteyn et al. (2014), Castillo et al. (2019, 2020), and Angerer et al. (2023), among others, have shown that, indeed, children's patience positively influences school outcomes, school-track choices, and labor-market outcomes. Similarly, Figlio et al. (2019) study immigrant children and how their home-country culture affects school outcomes. They also find that higher levels of long-term orientation are associated with better educational outcomes.

The impact of risk-taking on educational outcomes has first been conceptualized in Weiss (1972) and Levhari and Weiss (1974). Investments into human capital are inherently risky due to the uncertainty about one's ability, the parameters of the education production function, the quality of schools, and the later labor market conditions, among others. The relationship between the willingness to take risk and educational outcomes remains empirically unclear. While risk-lovingness might be beneficial for investment in higher education leading to jobs with higher earnings variance (Hartog and Diaz-Serrano, 2007; Hartog, Diaz-Serrano, et al., 2014), risk-aversion might also induce individuals to pursue higher education leading to jobs with lower risk of unemployment (Woessmann, 2016a). In fact, Brown et al. (2012) show that the additional investment of going to high school is more common among more risk-averse students while investments into college education are favored by higher risk-lovingness. On the other hand, Castillo et al. (2018, 2019) find that high-school *graduation* is negatively related to risk-taking preferences.

The main contribution of this dissertation is to show that these inter-temporal cultural traits can explain within- and across-country differences in student achievement. Chapter 2 com-

bines standardized test scores from the PISA study with newly available measures on personality traits from the Global Preference Survey. A simple cross-country regression shows that patience and risk-taking account for two-thirds of international achievement differences. Using migrant students and the traits from their cultural origin can help getting closer to a causal estimate of the relationship since this approach only compares migrants from different origin countries living in the same destination country. The migrant analysis confirms the findings from the cross-country setting and shows that patience is positively related to student assessment whereas risk-taking is negatively connected to skills. Most importantly, this chapter highlights the importance of considering the two traits jointly, which avoids underestimating the respective importance of each trait.

Chapter 3 builds on this study to explain differences in student achievement *within* countries. Understanding determinants of differences in student achievement within countries is potentially relevant for explaining regional differences in economic outcomes and these determinants are thus far understudied. Furthermore, the within-country setting allows for better identification of potential effects since it holds constant any country-level common factors that might bias the results. Chapter 3 uses Facebook marketing data and machine learning techniques to predict patience and risk-taking for regional entities. A validation exercise of the measures of patience and risk-taking with the countries used in the previous study confirms the suitability of these social-media data to predict cultural traits. Results from chapter 3 suggest that the social-media derived measures of patience and risk-taking not only account for achievement differences for a wider set of countries, but also for regional differences in student achievement in Italy and the US.

The findings from chapters 2 and 3 provide important insights to national and international differences in student achievement, but also highlight the importance of cultural differences between these (sub-)national entities that should be taken into account when designing education policies. While the studied inter-temporal preferences are malleable especially at the individual level (see e.g. Alan and Ertac (2018) and Jung et al. (2021)), cultural trails of larger entities such as countries tend to be rather stable over time (see e.g. Guiso et al. (2006); Bisin and Verdier (2011)). This implies that simply adopting policies from other countries or regions within a country will most likely not automatically lead to educational improvements if the implementation of these policies does not take into account the underlying cultural differences.

1.3 Feedback and Student Outcomes

Similarly to the more general cultural environment, the specific learning environment is a potential input factor for educational outcomes. In order to decide on education investments, students need to get a better understanding of their current educational success, e.g. through feedback provided by parents, teachers, and peers. In fact, feedback is part of any learning process as it aims to make visible where there is room for improvement of one's performance

and hence to reduce the uncertainty about the progress of one's skill production. For young individuals in education, this is an especially sensitive topic since their educational path coincides with the largest part of their personal development. Understanding how the way feedback is provided affects individual learning outcomes is hence crucial for educational success.

A series of studies has shown the beneficial effect of providing feedback in education (see e.g. Azmat and Iriberri (2010) and Goulas and Megalokonomou (2021)). Often times, these interventions consist in providing relative performance feedback to students that informs them about their position in the performance distribution relative to a certain peer group. This should help students by reducing the uncertainty about their level of performance through the additional information contained in the feedback. Furthermore, motivational effects of feedback have been highlighted in Muis et al. (2015) and Hermes et al. (2021).

This thesis contributes to this literature by studying a specific aspect of how to give feedback: the ordering of feedback elements. Chapter 4 shows evidence for the effect of feedback order on education outcomes from a field experiment with university students. In randomly varying order, these students received one positive and one negative feedback element on their performance in practice questions for a real exam. Participants who first got the positive feedback were more motivated to study for the exam compared to those who received negative feedback first. Students also adjusted their study content to the feedback topics. There was no average treatment effect of feedback ordering on exam performance, but the feedback ordering seemed to affect students depending on whether their negative-feedback topic was covered in the exam or not. Lastly, students showed emotional reactions to feedback ordering, i.e. those who got the positive feedback first stated to feel much better overall about the feedback they received.

The findings from chapter 4 point to mechanisms of feedback beyond its information content. Since the informativeness of the feedback did not differ by feedback ordering, the treatment effects on motivation and feelings about feedback imply an emotional reaction as suggested by the psychological literature (see e.g. Erickson et al. (2021)). These feelings can then influence (cognitive) behavior (see e.g. Zadra and Clore (2011), Baumeister et al. (2007), and Tyng et al. (2017)). The results from chapter 4 hence suggest to enrich the interpretation of feedback in economics by potential reactions of individuals to more than the information content of feedback.

1.4 Gender and Adult Skills

While the analysis of education and human capital is mostly focused on young individuals during (compulsory) schooling, learning is really a life-long process. Cognitive skills can change over time, and they are also highly related to individual outcomes, especially on the labor market. For example, higher cognitive skills are on average related to higher wages (Hanushek

et al., 2015). Although most (developed) countries have similar lengths of compulsory schooling, the resulting skills are distributed quite unequally between and within countries. This is most likely related to life choices individuals make based on e.g. their preferences, cultural norms, and the institutions of the country they live in.

One of the most widely discussed inequalities in educational and labor-market outcomes is the difference across genders. The related literature has almost exclusively studied individuals along the binary definition of gender and has found large differences between females and males. Such gender differences have for example been documented for cognitive skills of school children (Hyde et al., 2008; Contini et al., 2017). Boys tend to outperform girls in numeric skills that are known to be most predictive of later labor-market outcomes such as wages (Hanushek et al., 2015). These differences in favor of males are especially visible at the top of the skill distribution (Ellison and Swanson, 2010; Robinson and Lubienski, 2011; Contini et al., 2017) or even at both tails of the math distribution where boys are over-represented at both ends (Autor et al., 2020).

The contribution of this thesis is to focus on gender differences in cognitive skills among adults. Except for a very small number of empirical studies (Christl and Köppl-Turyna, 2020; Rebollo-Sanz and De la Rica, 2020), these have largely remained understudied which was in part due to a lack of suitable data. Chapter 5 exploits the availability of a new dataset to thoroughly study differences in labor-market-relevant numeracy skills between men and women across a wide range of countries. Although educational and labor-market outcomes for women and men have converged over the last decades, there still are large inequalities in numeracy skills favoring men. Furthermore, the analysis of these gaps along the wage distribution shows two main patterns: at the top of the wage distribution, women have both lower numeracy skills as well as lower returns to higher numeracy levels. These patterns are especially pronounced for parents and for individuals with their highest degree in a non-STEM field of study.[1]

The results from chapter 5 highlight the importance of cognitive skills in relation with gender gaps in other labor-market outcomes. Although the structure of the dataset does not allow for a causal analysis of the direction of this relationship, the findings suggest that skills and wages are highly interrelated and that life decisions such as children and the choice of a field of study can affect labor-market outcomes via cognitive skills. This has important implications for policies aiming at reducing gender inequalities in the labor market that should take into account both the accumulation and the depreciation process of skills needed on the labor market.

[1] STEM stands for *Science Technology Engineering Mathematics*. The enrolment into STEM or non-STEM fields of study is known to differ strongly by gender (see e.g. Goulas et al. (2022)).

1.5 Data and Empirical Methods

A sizeable part of this dissertation uses large-scale skill assessment data that provide internationally comparable measures of human capital. Chapters 2 and 3 make use of the so-called Programme for International Student Assessment (PISA), a standardized assessment of math, reading, and science skills conducted since 2000 among 15-year olds in almost 90 countries (OECD, 2019). Similarly, chapter 5 exploits the introduction of the Programme for the International Assessment of Adult Competencies (PIAAC) run in almost 40 countries among adults aged 16-65. It measures their skills in numeracy, literacy, and problem-solving in technology-rich environments that are 'necessary cognitive skills for advancing at work and participating in society'.[2] Both datasets do not only contain comparable skill measures that are particularly useful for (inter)national comparisons, but also offer a rich set of background characteristics about the individuals that allow for an in-depth analysis of their input factors to the education production function.

Chapters 2 and 3 additionally build on international data on cultural traits. The Global Preference Survey (GPS) offers scientifically validated measures on several preference parameters collected from representative samples in 76 countries in 2012 (Falk et al., 2018). Instead, the data from the Facebook Marketing Application Programming Interface (API) used in chapter 3 come in a less handy way and have to be transformed into measures of cultural traits with a series of machine learning techniques that are described more in detail in chapter 3. Being able to show the relationship between student achievement and inter-temporal traits with such different datasets lends additional credibility to the results.

In contrast to these large cross-country datasets, chapter 4 relies on an own data collection effort among university students in Munich, Germany. This procedure is obviously associated with a considerable amount of additional work compared to using the publicly available datasets described before, but also offers the possibility to tailor the data collection perfectly to the research question. The data presented in chapter 4 are the result of three online surveys collecting pre-treatment data, running the actual experiment, and eliciting post-treatment outcomes. It relies heavily on the measures used in Resnjanskij et al. (2021) as well as elicitation formats from Exley and Kessler (2022) and Falk et al. (2018).

The collection of own data then allows for a controlled randomization of individuals into the different treatment conditions. If randomization works properly, this should yield a causal estimate of the effect of feedback ordering on student outcomes. In fact, the students in the sample from chapter 4 seem to be balanced on the collected characteristics and additionally, results are robust to the inclusion of a large set of background characteristics as control variables in the regressions. To be able to identify a causal effect in such a clean way is a major advantage of experimental settings compared to the use of observational data.

[2] More information on the PIAAC data can be found here: `https://www.oecd.org/skills/piaac/`.

Causal identification in natural settings is much harder to achieve and necessarily relies on assumptions about the underlying populations. This is why, a first step to capture potential causal effects can be descriptive analyses hinting at correlations between certain measures. Chapter 5 uses such descriptive methods to grasp gender differences in cognitive skills. In particular, quantile regressions and decomposition methods adapted to quantile analyses are used to study in detail the distributional dimension of such gender gaps. Although they cannot establish causal links, these descriptive methods provide some very important insights to gender differences in cognitive skills, especially regarding a series of life decisions that individuals make. Future research might complement these analyses by studying settings that allow for causal interpretation of estimations.

Chapter 2 expands this descriptive approach to regression analyses with a series of controls that can address some first major concerns to identification. More specifically, estimating the relationship between cultural traits at the country level and individual student outcomes can be complicated by confounding factors on the country level that are both related to cultural traits as well as educational achievement. To address these concerns even further, the migrant analysis in this chapter exploits variation in cultural traits within countries from different countries of origin of migrant students. Similarly, chapter 3 looks at differences in student achievement across regions *within* countries, i.e. within the same general schooling system. While these approaches certainly cannot account for all potential confounders to a causal interpretation of the resulting estimates, they represent the closest approximation to causality in this setting.

2 Patience, Risk-Taking, and Human Capital Investment across Countries*

2.1 Introduction

Each release of international student assessment data such as the PISA test brings both professional and popular discussions of the causes of national differences in test scores. Such differences attract widespread attention not only because of the national ranking aspect but also because they provide indices of skills that are important for individual earnings (Hanushek et al., 2015, 2017a) and economic growth (Hanushek and Woessmann, 2012, 2015). Yet the underlying reasons for national differences in performance are not well understood. One often discussed but seldom analysed explanation involves cultural differences. This paper, relying on newly available measures of time and risk preferences across countries, establishes a clear case for linking skill investments to national preferences: International differences in student achievement are strongly related to international differences in patience and risk-taking.

Past research gives a mixed picture of the sources of test-score differences across countries (Hanushek and Woessmann, 2011; Woessmann, 2016b). Commonly available measures of educational resources such as aggregate spending, class size, and teacher characteristics explain little of existing score variation. By contrast, institutional features of school systems including test-based accountability, local autonomy, and private-school competition provide some explanation of score differences. Additionally, the role of parents and families is consistently strong, although highly variable across countries. Yet, the deeper structural determinants of international differences in societal choices of schooling inputs and in the productivity with which they are converted into educational outcomes remain poorly understood.

We focus on the potential role of differences in intertemporal preferences across societies as constituting fundamental determinants of student achievement differences. Our conceptual framework – developed in greater detail in Appendix A2.1 – combines the usually separated literatures about optimal human capital investment and about education production functions in order to highlight the central nature of preferences underlying intertemporal decision-making. Moreover, while investment decisions are generally viewed from the individual perspective, many decisions on educational inputs – in particular about resources and

* This chapter is co-authored with Eric A. Hanushek, Philipp Lergetporer, and Ludger Wößmann. It is based on the paper 'Patience, Risk-Taking, and Human Capital Investment across Countries' published in The Economic Journal, Volume 132, Issue 646, August 2022, Pages 2290–2307.

school institutions – are taken at the group level rather than the individual level, making it hard to disentangle impacts of individual preferences from group preferences.[1]

Two components of national preferences are central to the relative valuation of net payoffs in the present versus the future: time preferences (patience) and risk preferences (risk-taking). Human capital investment decisions take time to effectuate and even longer before any returns are realized. Just as the rewards for schooling investments require patience from the investor, national differences in patience may lead to national differences in educational outcomes.

The role of risk-taking is more ambiguous a priori. On the one hand, in line with the negative role of risk-taking stressed in the crime literature (e.g., Freeman (1999)), a preference for risk-taking may negatively impact the human-capital production process. For example, it may induce students not to complete required homework even though they take the risk of being caught and reprimanded by parents or teachers. An increased willingness to take risk may therefore favour misbehaviour, reduce effort in studying, and carry through to lower educational performance. On the other hand, consideration of various forms of school-completion and labour-market risks produces indeterminate predictions on how risk attitudes may affect human capital investment (Levhari and Weiss, 1974). For example, larger earnings variance in higher-educated occupations may give rise to a positive association between risk-taking and higher-education investment (e.g., Hartog, Diaz-Serrano, et al. (2014)), but lower unemployment risk (e.g., Woessmann (2016a)) may induce the opposite association.

Importantly, the intertemporal nature of human capital investment, its riskiness, and the inherent interrelatedness of the two preference components (Halevy, 2008; Andreoni and Sprenger, 2012) imply that one cannot consider the impact of patience without simultaneously considering risk-taking, and vice versa.

Our empirical investigation is facilitated by the recent innovations in international preference measurement in Falk et al. (2018). Their Global Preference Survey (GPS) employs experimental means to validate survey instruments that can be used to collect systematic data on international differences in several preference parameters.

We combine the GPS data with PISA data on the educational achievement of close to two million students observed in seven waves from 2000-2018 across 49 countries. These data allow us to estimate international education production functions at the student level that bring out how country differences in national preferences affect the skills acquired by students.

Our baseline analysis finds a strong and competing relationship between the two preference components and students' educational achievement. Patience has a strong positive and

[1] Following the literature (e.g., Guiso et al. (2006) and Alesina and Giuliano (2015)), we at times use 'culture' as shorthand for these group preferences. Obviously, however, culture is a very broad concept that has been given many different interpretations and goes far beyond the two intertemporal national preferences studied.

risk-taking a strong negative association with test scores. The substantial positive correlation between the two preference components implies that looking at them individually leads to consequential understatement of their respective importance.

Together, the two aggregate preference measures account for two-thirds of the variation in country average scores. Thus, a significant portion of the cross-country variation in student achievement may be closely related to fundamental differences in national preferences.

Consistent with a leading role of national cultures, the associations of the preference measures with individual achievement are much stronger for native students than for migrant students who moved into the school system from a different country. Moreover, the findings are stable across separate subjects (math, science, and reading) and subsamples (OECD and non-OECD).

To explore the causal structure of these cross-country associations, we focus on migrant students in the PISA data. Across 48 residence countries, we observe the country of origin of over 80,000 migrant students from 58 countries of origin with preference data. Following Figlio et al. (2019), we assign migrant students the preference values of their country of origin and study the performance of migrant children from different origin countries observed in the same residence country. We include fixed effects for each residence country to separate the effects of cultural factors from potentially correlated effects of the education systems, economies, or other common features of the residence country.

Students from home countries with an aggregate one standard deviation (s.d.) higher patience perform about 90 percent of a s.d. better in math (equivalent to the learning gains of roughly three years of schooling), whereas students from home countries with one s.d. higher risk-taking perform about 30 percent of a s.d. worse (equivalent to roughly one year of schooling). Consistent with an intergenerational persistence of home-country preferences, results are larger for migrant students who do not speak the language of their current residence country at home. While this migrant analysis cannot rule out all potential biases, our results are insensitive to different country samples, subjects, genders, alternative preference measures, definitions of the migrant population, different amounts of student test-taking effort, and several adjustments for the selectivity of migration – the most obvious threats to identification.

To investigate various channels through which national preferences might influence student achievement, we link them to the proximate inputs of the education production function in a final descriptive analysis. Patience is significantly positively correlated with family inputs, school inputs, and residual achievement differences (which likely combine productivity differences with unobserved inputs) across countries. Risk-taking is negatively correlated with family and residual inputs. Our results point to particularly important roles for family and residual inputs.

Our analysis of student achievement follows the recent literature investigating the influence of cultural factors on economic behaviour and outcomes (Guiso et al., 2006; Alesina and

Giuliano, 2015). With our migrant student analysis, we also contribute to this literatures' focus on intergenerational transmission (e.g., Bisin and Verdier (2011); Alesina and Giuliano (2014)). Past study of international student achievement has treated cultural factors largely as a source of possible bias in estimating the effects of proximate inputs in a cross-country setting (e.g., Hanushek and Woessmann (2011); Woessmann (2016b)). Here we show the value of directly addressing the potentially more fundamental role of some cultural traits as underlying causes of achievement differences in their own right, explaining largely unanalysed elements of the nature of societal human capital formation. The large effects of national preferences are in line with the role of unobserved parental characteristics that De Philippis and Rossi (2021) find in cross-country achievement differences.

One central conceptual feature is combining the two artificially separated strands of human capital literature: optimal investment decisions and the educational production process for skill development. The human capital investment literature following Mincer (1958), Becker (1964), Ben-Porath (1967), and others has measured human capital by individuals' years of schooling, equating skill development directly to the time costs of the investment. Human capital investments are portrayed as an individual intertemporal optimizing decision involving varying time commitments over the life cycle. For simplicity and tractability, this literature abstracts from any differences in skills obtained from time in school. The education production function literature on the other hand focuses on individuals' qualitative skill differences, generally looking at individuals with the same investment of school years but with different investment inputs (e.g., Hanushek (1986)). With some variations, the relevant skills are systematically related to inputs of the individual, the family, and the public through various aspects of schooling. These two lines of research are in essence looking at the same issue – how human capital investment decisions translate into differences in economically relevant skills. Treating these lines of research together yields clear insights into the deeper forces affecting skill differences of individuals and nations.

We also contribute to the literatures on time preferences (e.g., Sutter et al. (2013); Golsteyn et al. (2014); Figlio et al. (2019)), risk preferences (e.g., Levhari and Weiss (1974); Castillo et al. (2018)), and their interrelatedness (e.g., Halevy (2008); Andreoni and Sprenger (2012); Castillo et al. (2019, 2020)). Consistent with the associations of preferences with individual outcomes, our results show that patience and risk-taking have important effects on countries' human capital investment. At the country level, our analysis also relates to work on long-run comparative development (e.g., Galor and Özak (2016); Sunde et al. (2022)) and immigrants (e.g., Abramitzky and Boustan (2017)).

The next section describes the data. Section 2.3 develops the baseline estimates of the relationship of preferences and human capital across nations. Section 2.4 delves deeper into the causal structure using the analysis of migrants. Section 2.5 explores the association of patience and risk-taking with proximate input factors as possible channels. Section 2.6 concludes.

2.2 Data

Our analysis combines international data on student achievement (section 2.2.1) and on preferences (section 2.2.2). Details are found in Appendix A2.2.

2.2.1 The Programme for International Student Assessment (PISA)

The Organisation for Economic Co-operation and Development (OECD) has conducted the PISA test since 2000. PISA assesses achievement in math, science, and reading of random samples of 15-year-old students on a three-year cycle (OECD, 2019), providing repeated cross-sectional data representative in each country-by-wave cell. PISA also elicits background information on students and schools that we use as controls and as measures of channels.

Over the seven waves of PISA testing, 2000-2018, a total of 86 countries participated at least once (see Appendix Table A2.1 for details of all samples). Our baseline cross-country analysis considers the subset of 49 countries that are also covered by the GPS, using achievement data from a total of 1,992,276 students from 263 country-by-wave observations.

In our migrant analysis, we include migrant students in any residence country as long as PISA identifies the country of origin and home-country GPS data are available. (The entire 2000 PISA wave drops out because of missing information on students' country of origin). We observe 80,398 migrant students (and up to 145,506 in a wider definition) from 58 countries of origin located in 48 residence countries.

In the different parts of our analysis, we use data from a total of 86 countries, 71 of which participated in PISA and 64 of which have GPS data.

2.2.2 The Global Preference Survey (GPS)

The newly available Global Preference Survey (GPS) provides scientifically validated, high-quality data on several preference parameters collected from representative samples in 76 countries (Falk et al., 2018).[2] Using probability-based sampling, the GPS covers around 1,000 respondents in each country surveyed in 2012. We collapse the GPS data to the country level to construct one representative measure for each preference parameter per country. In total, we use GPS data from 64 countries – 49 countries in the baseline cross-country analysis and 58 as countries of origin in the migrant analysis.

[2] Because the GPS provides scientifically validated preference measures from representative samples for a large set of countries, it has important advantages, discussed in Appendix A2.2, over common alternative international datasets with proxies for national preferences such as the World Values Survey (WVS) and the Hofstede (1991) data. Correlations of our measures of intertemporal preferences with these alternatives and with the remaining GPS preferences are found in Appendix Table A2.3.

The GPS measures preferences in six domains: patience and risk-taking (the two preference components underlying intertemporal decision-making that are our main focus here) plus positive reciprocity, negative reciprocity, altruism, and trust. The underlying survey items were selected in an ex-ante validation exercise based on their ability to predict incentivized choices in a controlled laboratory setting. Patience and risk-taking are each measured by a combination of one qualitative survey question and one hypothetical choice scenario, which are then combined into a single preference measure using weights from the validation procedure.

Larger values of patience mean that the individual is more likely to accept deferred gratification. Larger values of risk-taking mean that the individual is more likely to take risky outcomes compared to certain outcomes. We z-standardize the GPS measure of each preference domain in our respective analytical sample and collapse standardized preference measures to the country level. Consistent with the interrelation emphasized in the behavioural literature, there is a strong positive correlation between patience and risk-taking in the GPS data of 0.358 at the country level (see Appendix Figure A2.1).

2.3 Patience, Risk-Taking, and Student Achievement across Countries

This section provides a description of the association of student achievement with patience and risk-taking across countries. It guides our analysis of the causal structure of the cross-country associations in section 2.4.

2.3.1 Empirical Model

Our empirical approach contrasts with most empirical investigations of educational production functions that include a long list of possible variables in order to soak up potential impacts of families, schools, institutions, and cultural traits. Being interested in more fundamental determinants of educational achievement across countries,[3] we employ a parsimonious specification of an education production function that models the output of education as centrally determined by national preferences:

$$T_{ict} = \beta_1 Patience_c + \beta_2 Risk_c + \alpha_1 \mathbf{B}_{ict} + \mu_t + \varepsilon_{ict} \tag{2.1}$$

[3] Moreover, to the extent that proximate inputs such as family inputs, school resources, and institutional features are themselves the outcomes of intertemporal choice decisions, they are bad controls in a model depicting the overall effect of national preferences on student achievement (see Appendix A2.1). Section 2.5 provides an analysis of these proximate inputs as potential channels of the impact of national preferences.

where achievement T of student i in country c at time t is a function of the two preference components of the country, a parsimonious vector of control variables $\mathbf{B}$ (student gender, age, and migration status), and an error term ε_{ict}. Fixed effects for test waves μ_t account for average changes over time along with any idiosyncrasies of the individual tests. Our coefficients of interest are β_1 and β_2 which characterize the relationship between the two preference components of a country's society – patience and risk-taking – and student achievement.

To account for the country-level nature of the main treatment variables, we cluster standard errors at the country level throughout. All regressions are weighted by students' sampling probabilities within countries and give equal weight to each country. In our analysis, original PISA scores are divided by 100 to convert achievement in into standard deviations.

2.3.2 Results of the Baseline Analysis

Results of the baseline model indicate important and intertwined roles of patience and risk-taking in international student achievement. Table 2.1 shows our baseline analysis of the association of student math achievement with patience and risk-taking across countries. When entered individually, there is a strong significant positive association of student achievement with patience (column 1) and a weaker, marginally significant negative association with risk-taking (column 2). Strikingly, both associations become much stronger (in absolute terms) and statistically highly significant when the two preference components are considered together (column 3), highlighting the importance of accounting for their interrelatedness. A one standard deviation (s.d.) increase in patience is associated with a 1.23 s.d. increase in student achievement, whereas the same increase in risk-taking is associated with a 1.24 s.d. decline in student achievement. Conditioning on the other component is particularly relevant for risk-taking: That part of the variation in risk-taking that is unrelated to patience has a strong negative association with student achievement.[4]

The results on patience and risk-taking are hardly affected when taking measures of other preference domains into account (column 4). In fact, none of the other four GPS measures – positive reciprocity, negative reciprocity, altruism, and trust – is quantitatively or statistically significantly associated with student achievement across countries. Thus, the preference components directly linked to intertemporal decision-making, rather than other preference domains, appear most relevant for educational achievement.

The interrelationship of the intertemporal preference components and achievement is depicted graphically in Figure 2.1. The upper panel shows simple bivariate scatterplots between average PISA math scores (pooled across waves) and the GPS measures of patience (left)

[4] Results are very similar for girls and boys, although the (absolute) estimate for risk-taking is slightly smaller for girls (columns 1 and 2 of Appendix Table A2.4). An interaction term between patience and risk-taking does not enter the model significantly (not shown).

and risk-taking (right) at the country level.[5] There is a strong positive association of student achievement with patience and a weaker and less precise negative one with risk-taking. At the country level, the R^2 of the underlying regressions suggest that patience alone accounts for 40.9 percent of the cross-country variance in achievement, whereas risk-taking alone accounts for only 6.2 percent. Both associations become much stronger and more precise when conditioning on the respective other preference component in the lower panel. The two preference components together account for two-thirds of the variance in average student achievement across countries (R^2 = 0.672). Interestingly, this is substantially larger than the sum of explained variance accounted by the two measures separately, underscoring the off-setting interplay of the two intertemporal preference components. The figures also show that the overall associations are not driven by any strong outliers.

If cultural traits are driving the achievement results, one would expect the residence-country culture to be less important for migrants whose parents are less steeped in that culture and whose exposure to the new culture is less. When we look separately at native students and migrants, we find a much stronger role of residence-country preferences for native students than for migrant students.[6] Among native students, a one s.d. increase in patience is associated with 1.30 s.d. higher achievement, and the same increase in risk-taking is associated with 1.32 s.d. lower achievement (column 5 of Table 2.1). By contrast, among students with a migrant background the association is much lower (0.70 s.d.) for patience and only marginally significant (at 0.37 s.d.) for risk-taking (column 6). Both differences are statistically significant.

The difference in results between students with and without migration background is in line with a leading role of cultural traits as deep determinants of student achievement rather than other unobserved schooling factors of a country. It also motivates our migrant analysis below that considers the cultural traits of the migrant students' countries of origin.

2.3.3 Robustness Analysis

One interpretational concern with low-stakes achievement tests such as PISA is that they might not only measure students' cognitive skills but also their effort on the test itself which in turn may depend on students' conscientiousness, intrinsic motivation, and other related skills (e.g., Borghans and Schils (2012); Akyol et al. (2021); Gneezy et al. (2019)). Among a number of measures of students' test-taking effort derived for the 2009 PISA wave, Zamarro et al. (2019) find that the extent of item nonresponse (the share of unanswered questions) in the student background questionnaire that follows the actual achievement test explains the largest share of cross-country variation in test scores. We construct this measure for all PISA waves to test whether the strong association of the intertemporal preferences with PISA achievement partly

[5] Results are almost identical when estimating the PISA scores as country fixed effects in equation (2.1) that includes control variables (but not patience and risk-taking; results available upon request).

[6] Students are classified as migrants if both parents were born abroad. The migrant analysis in section 2.4 shows that our findings are insensitive to alternative definitions of the migrant population.

reflects lower test-taking effort among less patient and more risk-taking students. Indeed, lower patience and higher risk-taking do significantly predict lower test-taking effort (higher item nonresponse on the background questionnaire) both at the individual and country level (not shown), validating a cultural component of test-taking effort.

While test-taking effort is relevant for overall test achievement, it does not alter the results for the two preference components. Individual students' item nonresponse rates on the background questionnaire negatively predict achievement on the math test (column 7 of Table 2.1). But the coefficients on patience and risk-taking hardly change. The same is true when we additionally control for average item nonresponse of the country (column 8). This is despite the fact that item nonresponse has substantial quantitative relevance: At the country level, the coefficient estimate suggests that going from the country with the lowest (0.010) to the highest (0.108) average item nonresponse decreases the average PISA score by 0.42 s.d. Thus, while test-taking effort appears relevant in low-stakes test taking, it does not alter conclusions about the more fundamental preference-achievement nexus considered here.

Additional robustness analyses described in Appendix A2.3 show that qualitative results are very similar for OECD and non-OECD countries, for achievement in science and reading, and when restricting the analysis to the first PISA wave after the GPS observations. The appendix also shows results for the alternative preference measures of WVS and Hofstede.

2.4 Exploration into Causality: Migrant Analysis

An obvious concern with the cross-country regressions is that a country's national preferences are likely correlated with other omitted country characteristics, such as legal or economic factors, that affect human capital investments. While some of the variation in these country factors may be the outcome of the national preferences and thus constitute channels rather than omitted variables, there may also be independent variation that happens to be associated with the national preference measures. For instance, a culture of patience might foster the economic development in a country more broadly, making it impossible to distinguish whether a positive association between patience and student achievement is due to patience per se or to better well-being. To address concerns about the causal interpretation of the baseline analysis, we explore an identification strategy that analyses cultural differences among migrants.

2.4.1 Empirical Model

If patience and risk-taking truly are cultural factors that affect educational investment decisions, migrants should retain some influence of the culture of their home countries. If we compare achievement across migrant children from home countries with different preferences who attend school in the same country of residence, we break the link between the cultural traits and elements of the schools, institutions, and environments of the country

of schooling – something that cannot be done for natives. Following similar applications in Carroll et al. (1994), Giuliano (2007), Fernández and Fogli (2009), and Figlio et al. (2019), we estimate regressions of the following form:

$$T_{ioct} = \delta_1 Patience_o + \delta_2 Risk_o + \gamma_1 \mathbf{B}_{ioct} + \theta_c \times \mu_t + \varepsilon_{ioct} \tag{2.2}$$

where T is achievement of migrant student i from country of origin o observed in residence country c at time t. $Patience_o$ and $Risk_o$ are the cultural traits measured in the country of origin.

The specification includes residence-country fixed effects θ_c to remove all common economic, institutional, and schooling factors for each residence country. We pool the data across residence countries but only use variation within each residence country and not cross-country variation to estimate the preference impacts. In fact, our specification controls for a full set of residence-country by wave fixed effects $\theta_c \times \mu_t$ which account for wave-specific differences across countries. Standard errors are clustered at the country-of-origin level.

We begin with a rather narrow definition of migrants, including only students with parents who are both born in a different country than the testing country. We assign first-generation migrant students their country of birth and second-generation migrant students the country of origin of their father. Across all PISA waves, there are 80,398 first- and second-generation migrants from 58 countries of origin with GPS data observed in 48 residence countries.

2.4.2 Results of the Migrant Analysis

The migrant analysis confirms the strong positive effect of patience on student achievement from the baseline analysis, as well as a significant negative effect of risk-taking, albeit of smaller magnitude compared to its baseline estimate and to the effect of patience. Table 2.2 reports the main regression results for the migrant analysis based on equation (2.2). All regressions include 180 fixed effects for each residence-country by wave cell and control variables. When entered separately, student achievement is significantly positively related to patience in the students' home country (column 1) and insignificantly positively to risk-taking (column 2). In line with the previous cross-country findings, the coefficient on patience increases and the coefficient on risk-taking turns significantly negative when both are entered together (column 3), underscoring the interrelated and competing nature of the two cultural traits. Students from home countries with one s.d. higher patience perform 0.93 s.d. better on the PISA math assessment, and students from home countries with one s.d. higher risk-taking perform 0.29 s.d. worse.[7]

[7] Results do not differ significantly between girls and boys (columns 3 and 4 of Appendix Table A2.4).

Column 4 additionally includes controls for the four other national preference components of the country of origin. These cultural controls do not significantly affect student achievement and leave the significant effects of the two intertemporal preference components intact. In fact, the coefficient on risk-taking increases (in absolute terms) to -0.45 in this specification.

In sum, the migrant analysis confirms the strong and positive effect of patience on student skill development documented in the descriptive cross-country analysis, even with the same overall magnitude. Similarly, it replicates the negative effect of risk-taking once we account for patience, though the effect size is smaller.[8] The migrant analysis rules out that the cross-country results are due to omitted residence-country variables. There is, of course, the possibility of remaining biases, some of which we address in the following robustness tests.

2.4.3 Robustness Analysis

To account for differences in students' test-taking effort, columns 5-7 of Table 2.2 control for individual and country-of-origin mean item nonresponse rates in the PISA student background questionnaires. Even though both enter significantly in explaining scores, the results on patience and risk-taking again hardly budge after controlling for these proxies for student effort.

Identification in the migrant analysis depends on the extent to which the national preferences of the country of origin provide a good proxy for the students' and families' actual preferences. A proxy for the extent to which families still hold their country of origin's influence is whether they still speak the language of their country of origin at home, rather than adopting the language of their new host country. The effects of the two home-country traits are 0.17 and 0.20 s.d. larger for those students who do not speak the residence-country language at home compared to those who do (columns 8 and 9), although the differences are shy of statistical significance. These results are consistent with an interpretation that the treatment variables in the migrant analysis do in fact capture the impact of cultural values of the countries of origin.

Appendix A2.4 shows that qualitative results are very similar for OECD and non-OECD countries, for achievement in science and reading, for first- and second-generation migrants and migrants of different ages of migration, and for alternative migrant definitions. The appendix also shows results for the alternative WVS and Hofstede preference measures.

Finally, we investigate whether several possible dimensions of selective migration pose a threat to identification in our migrant analysis. As a start, we note that neither economic conditions in the home country nor socio-economic differences in family background drive

[8] The differences in the point estimates on patience in the migrant analysis of Table 2.2 (columns 3 and 4) to the respective specifications in Table 2.1 are only marginally significant ($p < 0.1$) in the specification without the four other preference components and statistically insignificant ($p = 0.360$) in the specification with the other preference components. Both differences are statistically highly significant ($p < 0.001$) for risk-taking.

the estimates of national preferences (column 2 of Appendix Table A2.5). Another way to address potential bias from fundamental background differences is to include fixed effects for the origin continent of the migrant students. Column 1 of Table 2.3 shows that effects get slightly stronger when variation across continents of origin is removed. This analysis also indicates that results are not driven by geographic clustering of preferences by continent or by (exogenous) outstanding performance of any specific group such as students from Asia, Europe, or Latin America.

Migrants tend to be a selected subgroup from their countries of origin (e.g., Borjas (1987); Grogger and Hanson (2011)). Note that migrant selectivity that is the same across the different origin countries that send migrants to a specific residence country does not bias the migrant results. But differential migrant selectivity that is correlated with average cultural traits of the sending countries could introduce bias. This type of selection bias should be more severe for countries of origin with higher variance in cultural traits. However, the standard deviations of the two preference measures within the country of origin do not enter the model significantly and do not affect the qualitative results (column 2).

Another way to gauge the relevance of differentially selective migration is to take into account the geographical and cultural distance between sending and receiving countries. A general pattern in the migration literature is that migrants from neighbouring countries may be less positively selected than migrants from more distant countries (see Hanushek et al. (2017b), for US evidence), possibly because fewer hurdles have to be overcome. Controlling for the geographical distance between migrants' country of origin and residence country (using the distance measures from Mayer and Zignago (2011)) does not change our qualitative results (column 3). In column 5, we test whether effects vary with the cultural distance between the migrant students' country of origin and their residence country, as measured by the absolute difference in the preference measures between the two respective countries. The positive impact of patience does not vary with cultural distance, whereas the negative impact of risk-taking attenuates as cultural distance increases.

We also employ one direct measure of the differential selectivity of migrants based on their educational attainment. For each pair of sending and receiving countries, we compare the educational attainment of migrants in the residence country to the educational attainment of the populations of their respective countries of origin. We then measure migrant selectivity as the percentile of the country-of-origin distribution of educational attainment from which the average migrant in each residence country comes. Hanushek et al. (2017b) produced this measure for immigrants into the US, and we extend that analysis to the full matrix of origin and residence countries with available data. The measure of migrant selectivity is indeed positively associated with student achievement (column 7), but accounting for this differential selectivity does not affect our estimates of the impact of patience and risk-taking.

2.5 Channels of Impact

Our analysis has established robust relationships between the two preference components and student achievement without direct reference to underlying mechanisms. In the context of the canonical human capital production function, national preferences may influence student achievement through proximate inputs at the family, school, and institutional level as well as the productivity with which inputs are transformed into outcomes (see Appendix A2.1).[9]

To investigate the potential channels through which the national preferences operate, we regress four country-level variables reflecting major categories of proximate inputs[10] on the two preference components, patience and risk-taking (upper panel of Table 2.4).[11] Patience is positively associated with all four input components, although the association with institutional inputs is not quite significant at the 10 percent level. The association and explained variance are strongest for family inputs (column 1) followed by the residual (column 4). The residual factor has the character of total factor productivity, combining any unmeasured input components with the effectiveness of input use. Similarly, risk-taking is negatively correlated with all four input components, although only significantly so for family inputs and the residual.

As the estimation underlying the input aggregation may be biased by omission of the deeper preference variables, the presented estimates serve as an upper bound. A similar aggregation estimation including controls for the two national preferences can serve as a lower bound. The lower-bound procedure yields similar qualitative results of significant positive associations of patience with family and school inputs and a significant negative association of risk-taking with family inputs, only with expectedly smaller magnitudes (lower panel of Table 2.4). Interestingly, none of the other GPS preference measures (positive reciprocity, negative reciprocity, altruism, and trust) are significantly related to any of the input factors (not shown).

The observed patterns appear intuitive and highlight that the different proximate inputs – and particularly family inputs and residual productivity – may operate as channels through which the two intertemporal preferences affect student achievement. Of course, this analysis is inherently descriptive and should not be interpreted as a causal mediation analysis.

[9] Appendix A2.5 shows descriptive analysis that includes the proximate inputs as controls in our main analysis. Results indicate that a substantial part of the overall effects of the two preference components may work through the channels of these proximate inputs.

[10] Appendix A2.6 describes the construction of the four country-level input measures.

[11] Note that the analysis of channels is not meaningful for the migrant analysis. Migrants are not exposed to the school and institutional environment of the country that defines their cultural origin.

2.6 Conclusions

International differences in student achievement are at the forefront of many education policy debates, but the deeper reasons for why students in some countries perform better than in others are not well understood. While cultural differences have standardly been discussed as confounding factors in cross-country analyses of student achievement, we explicitly investigate specific cultural factors as deep determinants of student learning and skill investment. We focus on patience and risk-taking – the two preference components that reflect the intertemporal and risky nature of educational decisions – and combine international student achievement data from PISA with newly available data on national preferences from the Global Preference Survey.

In our cross-country analysis, patience is strongly positively and risk-taking negatively associated with student achievement. Importantly, ignoring the interrelatedness between the two positively correlated preference components leads to a substantial underestimation of both effects.

These results are confirmed in an identification strategy that compares migrant students from different country-of-origin cultures observed in the same residence country, eliminating any potential residence-country confounders. In a final descriptive analysis, we show that national preferences likely influence educational achievement by affecting several proximate inputs of the education production function, in particular family inputs and residual productivity.

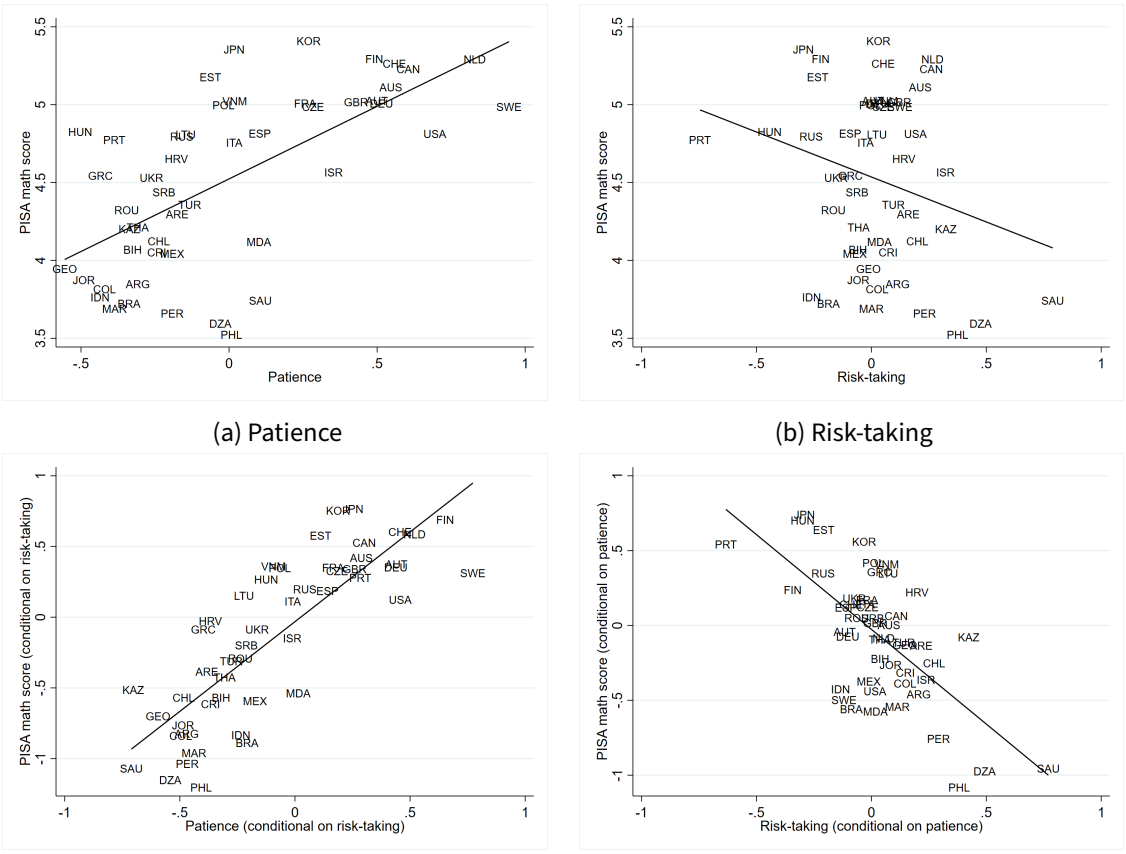
Taking an international perspective in studying the factors that influence student achievement comes with both advantages and challenges. The interest of this paper is understanding the relationship between national preferences and student achievement across countries, and the documented strength of the preference-achievement nexus indicates the first-order nature of this question. However, identifying causal effects in international data is particularly challenging because of the multitude of potential factors influencing student achievement. Our migrant analysis, together with a series of robustness analyses, are entirely consistent with the conclusions from the cross-country analysis. While addressing the most significant threats to identification of impacts of national preferences, other threats may remain. At the same time, it seems quite unlikely that any remaining bias would operate to eliminate the extraordinarily strong impacts of national preferences that we estimate.

While our results are important for understanding international achievement differences, they do not lend themselves to direct policy conclusions. Cultural traits of countries are slow moving and not easy to change (e.g., Guiso et al. (2006); Bisin and Verdier (2011)). At the same time, the relevant preferences are clearly amenable to change both at the individual and national level (e.g., Bird (2001); Alan and Ertac (2018); Jung et al. (2021)). The insight that cultural traits matter for educational achievement should thus be accounted for when designing policy interventions, particularly those focused on family inputs.

Our results imply that any policy intervention needs to take into account the fundamental role that cultural traits play in setting the context and in facilitating achievement. National policies cannot simply copy another country's experience. Failure to consider context may also explain why many previous attempts at international improvement have been unsuccessful. Finally, the finding that national preferences have limited association with institutional factors suggests that improving the institutional structures of school systems – whose importance has been highlighted by prior analyses (Hanushek and Woessmann, 2011; Woessmann, 2016b) – is a viable policy mechanism for improvement that does not necessarily depend on cultural change.

Figures and Tables

Figure 2.1 : Patience, Risk-taking, and Student Achievement across Countries



(a) Patience

(b) Risk-taking

(c) Patience (conditional on risk-taking)

(d) Risk-taking (conditional on patience)

Notes: PISA math score: average student achievement, 2000-2018. The added-variable plot in the lower left panel is created by first regressing both variables (math achievement and patience) on risk-taking. The residuals of the two regressions are then plotted against each other. These residuals represent the part of the variation in both variables that cannot be accounted for by risk-taking, assuring that risk-taking does not drive the depicted association. This exercise is numerically equivalent to regressing math achievement on patience and including risk-taking as a control variable. The equivalent procedure is used in the lower right panel. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table 2.1 : National Preferences and Student Achievement across Countries: The Intertwined Roles of Patience and Risk-taking

	Full sample				Natives	Migrants	Controls for test-taking effort	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Patience	0.917***		1.226***	1.186***	1.296***	0.698***	1.176***	1.117***
	(0.127)		(0.132)	(0.123)	(0.133)	(0.169)	(0.124)	(0.121)
Risk-taking		-0.482*	-1.241***	-1.314***	-1.320***	-0.371*	-1.200***	-1.141***
		(0.261)	(0.184)	(0.219)	(0.189)	(0.221)	(0.173)	(0.164)
Positive reciprocity				0.036				
				(0.226)				
Negative reciprocity				0.315*				
				(0.175)				
Altruism				-0.230				
				(0.188)				
Trust				-0.048				
				(0.152)				
Item nonresponse							-3.148***	-2.873***
							(0.158)	(0.151)
Item nonresponse (country mean)								-4.308***
								(1.046)
Control variables			Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,992,276	1,992,276	1,992,276	1,992,276	1,751,822	192,736	1,992,276	1,992,276
Countries	49	49	49	49	49	49	49	49
R^2	0.134	0.042	0.198	0.213	0.214	0.083	0.246	0.251
Difference between subsamples								
Patience					-0.598***			
					(0.149)			
Risk-taking					0.949***			
					(0.242)			

Notes: Dependent variable: PISA math test score in all PISA waves 2000-2018. Least squares regression weighted by students' sampling probability. Item nonresponse refers to the share of questions not answered in the student background questionnaire following the achievement test. Control variables: student gender, age, and migration status; imputation dummies; and wave fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table 2.2 : Patience, Risk-taking, and Student Achievement: Migrant Analysis

	Full sample							Language spoken at home	
					Controls for test-taking effort			Residence	Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Patience (country-of-origin)	0.779***		0.931***	1.032***	0.890***	1.021***	0.977***	0.718***	0.883***
	(0.115)		(0.116)	(0.133)	(0.114)	(0.100)	(0.105)	(0.117)	(0.151)
Risk-taking (country-of-origin)		0.183	-0.294**	-0.449***	-0.286**	-0.307**	-0.303**	-0.305**	-0.508***
		(0.210)	(0.122)	(0.140)	(0.119)	(0.120)	(0.114)	(0.115)	(0.165)
Positive reciprocity (country-of-origin)				-0.141					
				(0.157)					
Negative reciprocity (country-of-origin)				0.082					
				(0.087)					
Altruism (country-of-origin)				0.042					
				(0.144)					
Trust (country-of-origin)				-0.173					
				(0.138)					
Item nonresponse					-2.993***		-3.218***		
					(0.233)		(0.171)		
Item nonresponse (country-of-origin mean)							-3.319*		
							(1.691)		
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	80,398	80,398	80,398	80,398	80,398	36,668	36,668	48,556	24,520
Countries of origin	58	58	58	58	58	41	41	56	57
Residence countries	48	48	48	48	48	45		48	48
R^2	0.273	0.256	0.275	0.277	0.310	0.178		0.298	0.238
Difference between subsamples									
Patience (country-of-origin)								0.165	
								(0.130)	
Risk-taking (country-of-origin)								-0.203	
								(0.139)	

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions, including 180 fixed effects for each residence-country by wave cell. Sample: students with both parents not born in the country where the student attends school. Item nonresponse refers to the share of questions not answered in the student background questionnaire following the achievement test. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table 2.3 : Addressing Selectivity of Migrants in the Migrant Analysis

	Continent-of-origin fixed effects	Cultural variance	Migration distance			Selectivity of migrant schooling	
			Geographical	Cultural			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Patience (country-of-origin)	0.976***	0.818***	0.925***	0.933***	0.870***	0.987***	1.048***
	(0.125)	(0.191)	(0.117)	(0.149)	(0.137)	(0.105)	(0.106)
Risk-taking (country-of-origin)	-0.331**	-0.284**	-0.302**	-0.539***	-0. 903***	-0.300***	-0.411***
	(0.127)	(0.141)	(0.121)	(0.143)	(0.179)	(0.109)	(0.093)
Std. dev. of patience (country-of-origin)		0.285					
		(0.307)					
Std. dev. of risk-taking (country-of-origin)		-0.241					
		(0.372)					
Geographical distance (in 1000 km)			-0.010				
			(0.007)				
Patience (country-of-origin)					-0.287		
× Patience distance					(0.353)		
Risk-taking (country-of-origin)					1.048***		
× Risk-taking distance					(0.217)		
Selectivity of migrant schooling							1.269***
							(0.379)
Observations	80,398	80,398	80,398	29,019	29,019	39,725	39,725
Countries of origin	58	58	58	49	49	44	44
Residence countries	48	48	48	26	26	20	20
R^2	0.276	0.275	0.276	0.236	0.239	0.192	0.196

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions. Sample: students with both parents not born in the country where the student attends school. All specifications include residence-country by wave fixed effects and control variables (student gender, age, dummy for OECD country of origin, imputation dummies). Column-specific additional control variables: Col. 1: fixed effects for continent of origin. Col. 2: standard deviation of patience and risk-taking, respectively, in country of origin obtained from individual-level GPS data using individuals' sampling probability. Col. 3: geographical distance between respective residence and origin country according to most populous cities. Col. 5: interaction with difference in patience and risk-taking between respective residence and origin country (all variables demeaned). Col. 7: percentile of migrants' educational attainment on respective country-of-origin schooling distribution for each residence country. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; Falk et al. (2018).

Table 2.4 : The Association of Patience and Risk-taking with Proximate Inputs in the Education Production Function

	Family inputs (1)	School inputs (2)	Institutional inputs (3)	Residual (4)
Upper bound				
Patience	0.800***	0.069***	0.060	0.289***
	(0.087)	(0.021)	(0.037)	(0.095)
Risk-taking	-0.500***	-0.017	-0.066	-0.690***
	(0.139)	(0.033)	(0.059)	(0.151)
Observations	49	49	49	49
R^2	0.646	0.200	0.061	0.335
Lower bound				
Patience	0.382***	0.044**	-0.012	
	(0.062)	(0.019)	(0.027)	
Risk-taking	-0.325***	0.003	-0.009	
	(0.099)	(0.030)	(0.043)	
Observations	49	49	49	
R^2	0.461	0.120	0.008	

Notes: Country-level least squares regressions. Dependent variables indicated in column headers. Upper/lower bound refers to whether the preference variables are included in the underlying estimation of coefficients for the combination of the three input vectors. See text for details. Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Appendix

A2.1 Conceptual Framework

Our analysis of international differences in test scores is motivated by a desire to understand how systematic differences in national preferences contribute to variations in human capital across nations. This appendix provides a conceptual framework that discusses how national preferences related to intertemporal choices enter the human capital production model. We start by depicting educational choices in a human capital investment model with intertemporal preferences, incorporating several prior lines of inquiry into human capital investments (section A2.1.1). We then focus on the production of skills in order to understand how national preferences affect the individual and public choices of inputs into the education production function and the ultimate set of skills (section A2.1.2). Finally, we provide a deeper discussion of how patience and risk-taking enter separately and jointly into intertemporal decision-making (section A2.1.3).

Education as Intertemporal Choice

Educational decisions are fundamentally an intertemporal choice: initial investments of time, effort, and resources are set against expected future gains. Early human capital models thus directly related educational returns and investments to the rate at which future earnings are discounted (Mincer (1958, 1974); Becker (1964)). Further development of modern human capital theory naturally moved to optimal investment decisions of individuals, focusing on the maximization of lifetime earnings and stressing the time dimension of investments (Ben-Porath (1967) and Ben-Porath (1970); Heckman (1976); Rosen (1976)). The focus on a representative individual with perfect foresight precluded any deeper consideration of individual differences in intertemporal preferences. Given the intertemporal optimization decision, however, the two preference components related to balancing the present and the future – time and risk preferences – are crucial in understanding individual educational choices.

Surprisingly little explicit attention has been given to individual willingness to postpone gratification captured in patience, even though it is obviously a key element in the educational investment decision and thus in the earnings distribution. Detailed consideration of risk, by contrast, has entered human capital modeling at least since the contributions by Weiss (1972) and Levhari and Weiss (1974).[1]

The models of optimal human capital investment almost always focus on decisions about the quantity of education, which becomes the measure of individual skills. This focus has

[1] While the interaction of risk and human capital investment has been previously considered, attention has been confined mostly to labor-market outcomes. In considering different forms of labor-market risks such as variations in wages and employment, the optimal investment literature has produced varying predictions on how risk attitudes may affect human capital investment. Interestingly, because of its focus on labor-market outcomes, the analysis of risk in this literature has largely ignored how risk might also enter into the production process for skills. This separation of optimal individual investments from consideration of the underlying production process leads to considerable distortion in the analysis of the role of intertemporal preferences.

been natural given the availability of data and the consistency with the view of human capital investment as one of time. The perspective has been extraordinarily successful: The basic lifetime earnings model of Mincer (1974) has made years of schooling virtually synonymous with human capital in a wide range of empirical studies. Yet, school quantity is an imperfect measure of the underlying skill development that prescribes the optimality of downstream quantitative decisions in models of skill formation (Cunha et al., 2010) and that has future payoffs on the labor market (Hanushek and Woessmann, 2008).

Human Capital Investment, Educational Production, and National Preferences

Direct investigation of the production of skills has developed mostly separately from the study of optimal human capital investment (Hanushek, 1986). Research into skill development during the production stage focuses on what is actually learned, generally measured by achievement tests (rather than the time spent in school).[2] This research almost exclusively considers issues of technical efficiency of input usage and of the productivity of different inputs – without relation to human capital investment behavior.[3] But in reality, the education production function depicts how chosen inputs relate to human capital, as the observed proximate inputs to skill development are themselves the result of human capital investment decisions.

Further, even though the canonical human capital production model depicts skills as a function of family and school inputs, it is difficult to presume that these measured outcomes perfectly reflect the optimizing decisions of parents. The process of skill acquisition involves numerous actors – including the students themselves, their peers and friends, families, neighborhoods, teachers, school principals, and so on. Each presumably is optimizing over a different value function that may include different intertemporal preference parameters. Because of different assessments of the long-run value from human capital investments and different valuation of present versus future costs and payoffs, children may for example choose effort levels according to a preference for playing football or computer games over studying math in a way that diverges from what parents deem optimal in their maximization calculus.[4]

Importantly, many of the relevant educational investment decisions are actually made at the group level. How much to invest in school resources is usually publicly chosen at the municipal, state, or country level. Similarly, the institutional structures of school systems – features such as school accountability, autonomy, and choice which have been shown to matter greatly for

[2] Our focus on achievement scores does not imply a different interest from the school attainment work. We view the intermediate measures of adolescents' achievement as a good index of the ultimate skills of completed human capital investments. As an alternative, the Programme for International Assessment of Adult Competencies (PIAAC) measures the cognitive skills of adults, but analysis is hampered by limited country coverage.

[3] There are exceptions, for example, when the choices of parental investments are related to other inputs in the production function (Kim, 2001; Todd and Wolpin, 2003).

[4] Children may also be less willing or able to solve the dynamic optimization problem, leading to behavioral biases that prevent them from pursuing their own long-run well-being (Lavecchia et al., 2016).

student outcomes (Hanushek and Woessmann, 2011; Woessmann, 2016b) – are decided upon at the group level, and in most countries at the national level. As a consequence, aggregate societal intertemporal preferences will affect many parts of the education production process, making the set of preferences shared by the group important.

Therefore, in our analysis we change the perspective from individual preferences to group preferences and their relation to national cultures. Guiso et al. (2006), p. 23, define culture as "those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation." While different theoretical and empirical concepts and definitions exist (Alesina and Giuliano, 2015), relevant cultural values encompass the set of preferences shared by the group – including the intertemporal preferences that we deem important for educational choices.[5]

A key element of the existing cultural analyses is an emphasis on values that are transmitted persistently across generations (Bisin and Verdier, 2011; Alesina and Giuliano, 2014).[6] This persistent transmission motivates our empirical strategy below that looks at measures of national preferences in the home countries of migrants. Empirically, looking at migrants living in the same residence country allows us to distinguish cultural factors from other features of the residence country such as institutions and economies (Carroll et al., 1994; Giuliano, 2007; Fernández and Fogli, 2009; Figlio et al., 2019).

Recognition of intergenerational transmission also suggests some care in the specification of empirical models because common family factors may reflect cultural features. Analyses of educational production functions – whether within or across countries – commonly include measures of parental education (e.g., Hanushek (1986), Hanushek and Woessmann (2011)). But if national cultures influence human capital investment, parents' realized educational patterns may proxy the culture of their country. As such, they may partially be bad controls in the study of national preferences because they absorb part of the influence of the cultural factors.

More generally, a country's national preferences may affect all inputs in the education production process – on both the family and the school side – as well as the overall productivity with which these inputs are transformed into educational outcomes. This conceptualization implies that analyses of the effect of national preferences on student achievement should use very parsimonious specifications of the vector of control variables contained in the education production function.

[5] The concept of culture is related to the concepts of values, preferences, and personality traits and sometimes even subsumed in noncognitive skills, and the interrelations and distinctions between the concepts often remain vague. See Almlund et al. (2011) for an extensive discussion of the relationship between personality traits and preferences.

[6] Patience and risk-taking have been shown to correlate consistently between parents and their children at the individual level (e.g., Kosse and Pfeiffer (2012); Alan et al. (2017)).

Patience, Risk-Taking, and their Interrelatedness

While national preferences can encompass a wide variety of common traits, our interest in intertemporal decisions related to educational investments leads us to focus on two specific preference components: time preferences and risk preferences.

Time Preferences. The central role of the discount rate in models of optimal investment in human capital implies that time preferences are a key element of choices about whether to invest additional time, effort, and resources in improving educational outcomes. Preferences for payoffs in different time periods are reflected in patience, the trait of having a low rate of time discounting. For example, students must consider whether to give up play time with friends today – the opportunity cost of studying in the afternoon – for higher rewards in the future, such as graduating from school with better grades or the opportunity to receive better-paying jobs.[7]

It is remarkable that empirical studies only recently have begun to link validated measures of time preferences among students directly to educational outcomes. For example, Sutter et al. (2013) show that experimentally elicited measures of patience among Austrian children are significantly related to field behavior, including reduced violations of schools' code of conduct.[8] Using longitudinal Swedish data, Golsteyn et al. (2014) find that adolescents' time preferences are associated with human capital investments and lifetime outcomes. Castillo et al. (2019) show that experimentally elicited measures of discount rates among students in a school district in the U.S. State of Georgia are significantly related to high school graduation. Similarly, Castillo et al. (2020) show that preschool children in Chicago who are more patient are less likely to receive disciplinary referrals when they are in school years later. Combining the Hofstede (1991) cultural measure with migrant students in Florida schools as well as with the PISA data, Figlio et al. (2019) show that students from cultures with greater long-term orientation perform better on several measures of educational achievement.[9] At the macro level, Galor and Özak (2016) and Sunde et al. (2022) show that time preferences are importantly related to economic and educational outcomes in the long run.[10]

---

[7] As such, patience is closely related to similar concepts employed in the study of traits among children, such as the willingness to defer gratification as measured, e.g., by the famous "marshmallow test" (e.g., Mischel et al. (1989)), self-control (Moffitt et al., 2011), or perseverance and grit (e.g., Duckworth et al. (2007)).

[8] Alan and Ertac (2018) and Alan et al. (2019) show that measures of patience and grit are malleable to classroom interventions.

[9] Mendez (2015) shows the potential relevance of cultural traits for student achievement using a principal component from eleven different value questions in the World Values Survey (WVS) with migrant students in seven host countries in PISA, but the approach does not delve into specific components of national preferences. Cordero et al. (2018) include WVS measures in efficiency measurement of school systems in PISA.

[10] While formulated from a different perspective, a recent literature suggests that student behavioral differences related to effort, care, motivation, and perseverance may impact country test scores (e.g., Borghans and Schils (2012); Balart et al. (2018); Akyol et al. (2021); Gneezy et al. (2019); Zamarro et al. (2019)). These behavioral differences may in turn reflect underlying cultural differences. We return to the role of student test-taking effort in robustness analyses below.

Risk Preferences. Beginning with the empirical study of occupational choices by Weiss (1972) and the theoretical analysis in Levhari and Weiss (1974), a stream of studies of human capital investments explicitly introduced various components of uncertainty and risk. In a very general way, Levhari and Weiss (1974) consider a range of risky elements in labor-market investment decisions including future supply and demand conditions as well as knowledge of one's own ability, of how time and money convert into human capital, and of the quality of schools along the investment path. They show that it is not possible a priori to determine how risk affects human capital investment incentives, a conclusion reiterated in the extensive review by Benzoni and Chyruk (2015). For example, higher earnings variance in higher-educated jobs may give rise to a positive association between risk-taking preferences and investment in higher education (e.g., Hartog and Diaz-Serrano (2007) and Hartog, Diaz-Serrano, et al. (2014)), whereas lower unemployment risk of higher-educated jobs (e.g., Woessmann (2016a)) may give rise to the opposite association. In contrast to the indeterminate nature of the impact of different forms of labor-market risks, the role of risk-taking is more clear-cut when directly considering student behavior in the human-capital production process: Drawing on insights from the economics of crime, Castillo et al. (2019) argue that risk-lovingness may deter educational effort by favoring misbehavior in adolescence if there is uncertainty about getting caught by teachers or parents.

Existing empirical evidence on the association between risk and human capital investment is closely related to the specific components of risk considered in individual studies. Using U.S. data, Brown et al. (2012) suggest that lower risk-taking leads to more investment in high-school education compared to less than high school but less investment in college compared to high school. Analyzing both wage and employment uncertainty, Groot and Oosterbeek (1992) find different results on returns by type of schooling (vocational or college) in the U.S. and the Netherlands, while Koerselman and Uusitalo (2014) find little differential effect of lifetime income variability on different schooling choices in Finland. Palacios-Huerta (2003) compares human capital risks to financial assets risk and detects wide variation in risk-adjusted rates of return. Using direct measures of children's risk preferences, Sutter et al. (2013) find little evidence of associations with field behavior. By contrast, Castillo et al. (2018, 2019) show a negative association of risk-taking preferences with high-school graduation and a positive with disciplinary referrals, in line with the notion that lower risk-taking may keep students out of trouble during the human-capital production process.[11]

The Interrelatedness of Time and Risk Preferences. While much of the prior literature has considered time and risk preferences separately, behavioral economics has emphasized their inherent interrelatedness: since only the present can be certain and the future always

[11] There is also evidence of associations of patience and risk with intelligence among adults, again with mixed evidence on risk (Dohmen et al., 2010, 2018; Potrafke, 2019).

contains an element of uncertainty, it is inescapable that the two preference components are intertwined (Halevy, 2008; Andreoni and Sprenger, 2012).[12]

An important implication of this interrelatedness is the need to control for the one preference component when studying the effect of the other. In fact, because many of the studies of risk-taking do not control for patience, this interrelationship may help explain the reasons for the divergent empirical effects of risk on investment. Even more, given the a priori indeterminate direction of the effect of risk-taking in the human capital production function, the direction of bias when estimating the effect of patience without considering risk-taking is also unclear.

A2.2 Additional Information on the Data

The information on the PISA and GPS datasets provided in this appendix complements the basic information contained in section 2.2 of the main text.

The Programme for International Student Assessment (PISA)

The target population of the representative random sampling of PISA are 15-year-old students, independent of grade level or educational track attended (OECD, 2019). The sampling in most countries proceeds in two steps. First, a random sample of schools that teach 15-year-old students is drawn using sampling probabilities that assure representativeness. Second, 35 students aged 15 years are randomly sampled in each school.[13] PISA only reports data for countries that meet the OECD's high sampling and data-quality standards.

To create comprehensive measures of students' competencies, PISA has students complete a broad array of tasks of varying difficulty in assessments that last for up to two hours. The testing mode was paper and pencil until 2012 and changed to computer-based testing in 2015. PISA achievement in math, science, and reading were standardized to a mean of 500 test-score points and a standard deviation of 100 test-score points for OECD-country students in wave 2000 (and rescaled on the same metric again in 2003 in math and in 2006 in science). We divide PISA scores by 100 throughout to express achievement in percent of a standard deviation. As a rule of thumb for interpreting PISA scores, about a quarter to a third of a standard deviation corresponds to the learning gains of one year of schooling (Woessmann, 2016b). Table A2.2 shows descriptive statistics of country-level PISA achievement in the three subjects.

In addition to achievement data, PISA elicits background information on student and family characteristics using student questionnaires, as well as contextual information on school resources and the institutional environment using school questionnaires completed by school

[12] Their particular formulation has been questioned, but the basic concept seems clear. See the exchange in Cheung (2015), Epper and Fehr-Duda (2015), Miao and Zhong (2015), and Andreoni and Sprenger (2015).

[13] We use the first plausible value (PV) provided by PISA throughout. Our results hold when considering other PVs and when employing estimation procedures that explicitly account for imputation and stratified sampling in the PISA data (results available upon request).

principals. From these rich background data, we select core control variables – student gender, age, and migration status (first and second generation) – for our regression analysis. In addition, we derive measures of proximate inputs at the family, school, and institutional level that we use in our channel analysis. At the student level, these are parental education (six categories), parental occupation (four categories), books at home (four categories), computer for school work at home (dummy), and other language than the test language spoken at home (dummy). At the school level, these are school location (three categories), school size, share of fully certified teachers, and shortage of educational material (dummy). At the country level, these are GDP per capita, share of privately managed schools, share of government funding of schools, central exit exams (dummy), and a school-autonomy index. The share of missing values for these variables is generally very low, averaging 5 percent. We impute missing values using the respective country-by-wave mean and include imputation indicators (one dummy per variable that equals one if the respective variable is missing and zero otherwise) in our regression analysis.[14]

The complex structure of country inclusion in our analyses is explained in greater detail in the note to Table A2.1. In the migrant analysis, countries can be included as residence countries even if there are no GPS measures for them (as long as they participated in PISA and have migrant children from countries of origin that participated in the GPS) and as countries of origin even if there are no PISA measures for them (as long as they participated in GPS and have 'sent' students as migrants to PISA-participating countries).

The Global Preference Survey (GPS)

The Global Preference Survey (GPS) was conducted within the framework of the 2012 wave of the international Gallup World Poll, an annual survey on social and economic topics. Altogether, the GPS uses twelve survey items to measure preferences in the six domains. In an ex-ante validation exercise, students at the University of Bonn took different incentivized decisions in a controlled laboratory setting and answered numerous survey questions for each preference domain (Falk et al., 2022). The survey items were then selected based on their ability to predict the incentivized choices. For most preference domains, this exercise led to the selection of a combination of one qualitative survey question and one hypothetical choice scenario (see Falk et al. (2018) for details).[15] For each domain, the selected survey items are then combined into a single preference measure using weights from the validation procedure.

For patience, the qualitative survey item, elicited on an 11-point Likert scale, is: *'How willing are you to give up something that is beneficial for you today in order to benefit more from that in*

[14] In the few cases where a variable is missing for an entire wave in a given country, we impute by averaging over the country's other PISA waves. Dropping these country-by-wave observations as a robustness check does not affect our results (results available upon request).

[15] Exceptions are trust and negative reciprocity, which are measured using one and three qualitative survey questions, respectively. While the qualitative items elicited on Likert scales may be subject to reference bias in the cross-country setting, this is less likely for the choice scenarios.

the future?' The hypothetical choice scenario for patience entails a series of binary decisions between 100 Euro today or a higher amount in the future: *'Suppose you were given the choice between receiving a payment today or a payment in 12 months. We will now present to you five situations. The payment today is the same in each of these situations. The payment in 12 months is different in every situation. For each of these situations we would like to know which one you would choose. Please assume there is no inflation, i.e., future prices are the same as today's prices. Please consider the following: Would you rather receive amount x today or y in 12 months?'*

For risk-taking, the qualitative 11-point-scale question is: *'In general, how willing are you to take risks?'* The quantitative staircase measure is based on the question: *'Please imagine the following situation. You can choose between a sure payment of a particular amount of money, or a draw, where you would have an equal chance of getting amount x or getting nothing. We will present to you five different situations. What would you prefer: a draw with a 50% chance of receiving amount x, and the same 50% chance of receiving nothing, or the amount of y as a sure payment?'*

The GPS dataset does not provide responses to individual items, so we use the available combined preference measures in our analysis. The GPS dataset contains one z-standardized variable for each preference domain. Standardization is conducted at the individual level so that each preference has mean zero and standard deviation one in the individual-level world sample. For the purpose of our analysis, we z-standardize each individual preference measure in our respective analytical sample and collapse standardized preference measures at the country level. Table A2.2 presents descriptive statistics of the resulting data.

In the 49-country sample of our baseline analysis, there is a significant cross-country correlation between patience and risk-taking of 0.358 (depicted in Figure A2.1). Table A2.3 shows country-level correlations for all preference measures. While patience is not significantly correlated with the other four GPS preference domains (although there are marginal correlations with negative reciprocity and trust), there is a significant correlation of risk-taking with negative reciprocity.

The GPS has several important advantages over alternative international datasets with proxies for national preferences, because it provides scientifically validated measures of the two preference components underlying intertemporal decision-making from representative samples for a large set of countries. The closest alternatives are the World Values Survey (WVS) and the Hofstede (1991) data, both of which provide survey data on attitudes, beliefs, and personality traits across countries. While the WVS is based on representative samples, the Hofstede data are mainly based on IBM employees and are not representative. Importantly, in contrast to the GPS, the validity of the proxies for patience, risk-taking, and other preference domains from these surveys is unknown.[16] Reinforcing the high quality of the GPS data, Falk et al. (2018)

[16] For instance, the best proxy for patience in the WVS is an item on 'long-term orientation' that asks 'Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially

show that the GPS patience measure is more predictive of comparative economic development than the measures of long-term orientation from the other two datasets. Interestingly, the correlations of the GPS measures of patience and risk-taking with their respective proxies in the WVS and Hofstede datasets are limited: The correlations of GPS patience with the WVS and Hofstede long-term orientation measures are -0.060 and 0.247, respectively, and statistically insignificant (Table A2.3). The correlations of GPS risk-taking with WVS risk-taking and Hofstede uncertainty avoidance are only slightly stronger at 0.239 and -0.302, respectively. For our robustness analyses, however, we investigate the WVS and Hofstede data as alternative measures for patience and risk-taking (see Appendix A2.3 and Appendix A2.4).

A2.3 Additional Robustness Analyses for the Baseline Analysis

This appendix provides additional robustness analyses for the baseline analysis presented in section 2.3 of the main text.

Restriction to wave 2015 of PISA. Our main analysis pools the achievement data of all seven PISA waves (2000-2018), which is justified because cultural aspects by definition are focused on traits that are fairly unchanged over the long run. Moreover, the vast majority of country variation in PISA scores is between countries rather than over time. Pooling extends the country sample and provides more precise measures of long-run educational achievement. Table A2.6 shows that results are qualitatively the same when restricting the analysis to the 2015 PISA wave (the first PISA wave after the elicitation of the GPS data in 2012), indicating that the pooled analysis is not affected by the relative timing of the observation of preference and achievement data.

Country subsamples. To test whether our main results differ by level of development, columns 1 and 2 of Table A2.7 present separate regressions for OECD countries and non-OECD countries, respectively (measured as ever belonged to OECD). The qualitative pattern of our findings is very similar and does not differ significantly between the two subsamples.

In additional subsample analyses, we re-estimated the models in Table 2.1 (columns 3 and 4) excluding one wave or one country at a time. Qualitative results are insensitive to this alteration. The coefficients on patience and risk-taking remain significant in all these regressions (not shown).

Additional subjects. Our main analysis focuses on math achievement, which is generally conceived to be most readily comparable across countries compared to other subjects such

important?' and is coded 1 if the respondent selects the item 'thrift, saving money and things', and 0 otherwise. The Hofstede dataset contains proxies for long-term orientation and uncertainty avoidance that are composed of a collection of four qualitative survey items each, several of which appear somewhat unrelated to the concepts that they mean to measure. For example, long-term orientation includes an item on 'How proud are you to be a citizen of your country?' and uncertainty avoidance includes an item on 'All in all, how would you describe your health these days?' (see footnote 7 in Falk et al. (2018) for details).

as reading (which by construction is to some extent language specific). Reassuringly, results are very similar for achievement in science and reading (columns 1 and 2 of Table A2.8). In science, a one s.d. increase in patience (risk-taking) is associated with a test-score increase (decrease) by 1.12 s.d. (1.17 s.d.). In reading, the corresponding coefficient is 1.11 s.d. (1.13 s.d.). Thus, the reported associations are universal in the sense that they do not depend on a particular subject.

Alternative preference measures. Given the rather vague measurement of the underlying intertemporal concepts in the WVS and Hofstede datasets (see Appendix A2.2), we are less confident about the validity of these alternative measures. Still, Table A2.9 shows that the WVS cultural measures yield a similar pattern of a positive association of student achievement with long-term orientation and a negative association with risk-taking (column 1). Using the Hofstede measures, long-term orientation is significantly positively associated with student achievement, whereas uncertainty avoidance is insignificant (column 2).[17]

Oster (2019) analysis. We also perform an analysis of unobservable selection and coefficient stability following Oster (2019). We compare our baseline model (column 3 of Table 2.1) to a restricted model without the control variables and follow Oster (2019) in setting $R_{max} = 1.3\tilde{R}$. Results in column 1 of Table A2.10 indicate that, assuming $\delta = 1$, the estimated bias-adjusted treatment effect β^* for patience is 1.358, which is even larger than our baseline estimate because adding the controls to the restricted model increases the coefficient estimate. For risk-taking, β^* is -1.181, only slightly below our baseline estimate. Thus, both estimates remain substantial in the bounding analysis that assumes that selection on unobservables is as strong as selection on observables. In both cases, the value of δ for which $\beta = 0$ far exceeds the suggested cutoff of $\delta = 1$: for patience, $\delta = -18.093$, and for risk-taking, $\delta = 8.224$. That is, the degree of selection on unobservables would have to be several times as large as selection on observables to eliminate the main result.

Accounting for uncertainty in GPS estimates. The GPS measures of patience and risk-taking are effectively generated estimators, reflecting estimated sample means based on random samples of around 1,000 respondents in each country (see section 2.2.2 in the main text). This uncertainty in the generated regressors should be accounted for in the regressions (e.g., Pagan (1984); Murphy and Topel (1985)). We use a two-step bootstrap procedure where the first step draws a random sample of 992 observations with replacement in each country (the smallest number of observations within a country in the GPS data) and computes the sample means of patience and risk-taking for each country. The second step uses these preference values to run our main regression (column 3 of Table 2.1). Repeating this procedure 1,000 times, we calculate our coefficient estimates as the mean of these repeated estimations and the bootstrapped standard errors of our coefficient estimates as the standard deviation in

[17] In a specification that includes all preference measures from the GPS, WVS, and Hofstede together, only the GPS measures of patience and risk-taking remain large and statistically significant, whereas the WVS and Hofstede measures lose their statistical significance (not shown). The GPS results are also robust to controlling for WVS trust, which does not enter significantly (not shown).

the sample of 1,000 estimated coefficients. The bootstrapped coefficients on patience and risk-taking are very similar to the ones reported in the paper and remain statistically highly significant (results available upon request).

Uncertainty in the earnings returns to education. To scrutinize the relative importance of different potential channels of the relationship between risk-taking and human capital investment discussed in Appendix A2.1, we conducted additional regression analyses in which we account for the uncertainty in earnings returns to education (proxied by the R^2 of Mincer-type regressions; see Hanushek et al. (2015, 2017b)). The results show that the coefficient on risk-taking becomes less negative as earnings uncertainty increases. At face value, this finding suggests that both (i) a negative effect of risk-taking suggested by the crime literature and – to a lesser extent – (ii) a positive effect of risk-taking suggested by studies highlighting the earnings variance in higher-educated occupations may act simultaneously on student achievement (not shown).

A2.4 Appendix: Additional Robustness Analyses for the Migrant Analysis

This appendix provides additional robustness analyses for the migrant analysis presented in section 2.4 of the main text.

The regressions of the migrant analysis include 180 fixed effects for each residence-country by wave cell. Table A2.11 shows that results are very similar in specifications that include 48 fixed effects for the respective residence countries and six fixed effects for waves, but not their interactions.

Country subsamples. Results of the migrant analysis also do not differ significantly by the level of development of migrants' countries of origin. Patience enters significantly positively, and risk-taking significantly negatively, in the subsamples of migrant students from both OECD and non-OECD countries of origin (columns 3 and 4 of Table A2.7). The positive point estimate of patience is somewhat larger in OECD countries, whereas the negative point estimate of risk-taking is somewhat larger (in absolute terms) in non-OECD countries. However, neither difference is statistically significant.

Additional subjects. Columns 3 and 4 of Table A2.8 present results for student achievement in science and reading, respectively. Results are again very similar to math, although the negative coefficient on risk-taking is not statistically significant in the other two subjects.

Alternative preference measures. The qualitative pattern of results on patience is also confirmed with the alternative measures in the WVS and Hofstede datasets (columns 3 and 4 of Table A2.9). In both cases, there is a significant positive effect of long-term orientation on student achievement, in line with the results in Figlio et al. (2019). The WVS data also confirm a significant negative effect of risk-taking. By contrast, the Hofstede risk measure points in the opposite direction – a negative effect of uncertainty avoidance – which presumably reflects

the poor measurement of the underlying concept by the items contained in this proxy (see Appendix A2.2).

Different migrant subgroups. In addition to distinguishing by the language spoken at home (see section 2.4.3 in the main text), another migrant subgroup analysis distinguishes migrants by the time at which the students themselves migrated to the country of residence. Results in columns 1 and 2 of Table A2.12 show that the effect of patience does not differ significantly between second-generation migrants (born in the residence country after their parents had migrated) and first-generation migrants (born in the country of origin), and the negative effect of risk-taking is actually larger for second-generation migrants.

We can exploit information on the age of migration in our dataset to subdivide the first-generation migrants further by whether they arrived in the residence country before or after age 6, when they would usually start school. Within the first-generation migrants, the effects of the two preference components do not differ significantly by whether students had migrated earlier or later (columns 3 and 4). While these patterns show the robustness of our main findings, they do not support the notion that later migrants hold onto more of their country-of-origin culture.

Alternative migrant definitions. Our main specification uses the country of origin of the students' fathers for reference for second-generation migrant students. Results in the first column of Table A2.13 show that estimates are virtually identical when the country of origin of the mother is used instead. Column 2 uses the average value of the national preferences of the country of origin of both parents when both are available, and the measure of the respective country of origin of the father or mother if the information is available only for one of them. Again, results hardly change in the slightly larger sample of 83,798 students.

As the PISA data allow us to observe both parents' country of origin, we can also enter the national preferences of fathers' and mothers' country of origin simultaneously (column 3). While this horse-race specification is identified only from children whose parents come from different countries, results still provide a relatively clear pattern. The effect of patience is significantly positive for both parents, although it is twice as large for fathers' compared to mothers' patience. By contrast, the effect of risk-taking is fully driven by fathers, with risk-taking of the mothers' country of origin not entering migrant students' achievement.

In our main specification, we adopt a rather narrow definition of migrants that includes only students whose parents are both born in a different country than the testing country. Alternatively, we can use a wider definition that includes all students with at least one parent born abroad – defining as natives only those with both parents born in the testing country. This wider definition increases the number of observations to over 140,000 migrant students. While, expectedly, point estimates are slightly smaller with this broader measurement, results are in fact very similar to those in the smaller sample with the narrower definition, independent

of whether the country of origin is defined based on the mother, the father, or the average (columns 4-6).

Including the national preferences of both parents' countries of origin simultaneously also again yields very similar results (column 7).

In a few cases, the effect of a specific country of origin is identified from only a limited number of student observations, potentially introducing substantial measurement error for these countries of origin. However, if we restrict the analysis to cases where at least 50 students are observed from each country of origin – which reduces the number of countries of origin from 58 to 46 – results remain virtually unaffected (column 8).[18]

Selective migration. In addition to the detailed analysis of selective migration presented in the main text, we also estimated a specification that interacts the preference variables with the raw difference in the preference variables between country of origin and residence country (rather than the absolute difference as in column 5 of Table 2.3). None of the interactions is significant, indicating that effects do not differ by whether migrants go to countries with higher versus lower preference values (not shown). The results remain unchanged when the controls for geographical and cultural distance are entered together (not shown).

Oster (2019) analysis. An analysis of unobservable selection and coefficient stability following Oster (2019) again suggests that results remain stable. Comparing our baseline model (column 3 of Table 2.2) to a restricted model without the control variables and setting $R_{max} = 1.3\tilde{R}$, results in column 2 of Table A2.10 indicate that the estimated bias-adjusted treatment effect β^* (assuming $\delta = 1$) is 0.875 for patience and -0.264 for risk-taking, very close to our baseline estimates. Accordingly, selection on unobservables would have to be 13.6 the selection on observables for patience and 9.8 for risk-taking in order to eliminate the main result.

Accounting for uncertainty in GPS estimates. To account for uncertainty in the generated regressors, we again implement the two-step bootstrap procedure described in Appendix A2.3. Again, point estimates are very similar to our baseline model and estimates remain statistically highly significant (results available upon request).

A2.5 Models with Extended Controls

Estimates of education production functions usually contain measures for proximate inputs – family inputs, school resources, and institutional features. To the extent that these proximate inputs are themselves the outcomes of intertemporal choice decisions, they would be bad controls in a model depicting the overall effect of national preferences on student achievement. Including proximate input factors in our model, however, provides a descrip-

[18] This robustness analysis drops the following countries of origin: Bangladesh (17 observations), Canada (1), Chile (47), Finland (2), Georgia (3), Indonesia (27), Kazakhstan (34), Lithuania (3), Moldova (11), Nigeria (4), Saudi Arabia (8), and Thailand (20).

tive evaluation of the importance of these input channels and shows the robustness of the preference-achievement association to consideration of variation in input factors that stem from other sources. Therefore, in this appendix we report specifications that include a rich set of proximate input factors as control variables that would generally be included in education production functions:

$$T_{ict} = \beta_1 Patience_c + \beta_2 Risk_c + \alpha_1 B_{ict} + \alpha_2 F_{ict} + \alpha_3 S_{ict} + \alpha_4 I_{ct} + \mu_t + \varepsilon_{ict}$$

which, in addition to our baseline model (2.1), includes measures of the inputs from student's families F, schools S, and institutional structures of school systems I.

When adding the proximate inputs as control variables to the model of the baseline analysis, the coefficient estimates on the two preference components remain large and statistically highly significant, but are reduced in size (column 1 of Table A2.5). The extended set of controls for family, school, and institutional inputs (described in the table notes) are likely bad controls because they too are outcomes of the deeper cultural traits. The reduction of the coefficients on patience by 39 percent and on risk-taking by 33 percent (when comparing column 3 of Table 2.1 to column 1 of Table A2.5) in this descriptive analysis indicates that a substantial part of the overall effects of the two preference components may work through the channels of these proximate inputs.

Column 2 of Table A2.5 provides equivalent results for the migrant analysis. The specification adds the set of extended controls on family and school inputs in the residence country, as well as the country of origin's GDP per capita. This latter control addresses the concern that, for instance, better performance of migrants from high-patience countries merely reflects differences in income (as opposed to genuine effects of cultural traits). As expected, the coefficient on patience is reduced in this specification (because the family and GDP controls may take out some of the total effect of cultural traits), but it remains large and significant (as does the coefficient on risk-taking).[19]

Section 2.5 in the main text provides a closer analysis of the association of the different input factors with the two preference components.

A2.6 Details of the Channel Analysis

In order to shed light on potential mechanisms, section 2.5 in the main text considers how patience and risk-taking relate to the major categories of proximate inputs – family inputs, school inputs, and institutional inputs. The starting point of the channel analysis is devel-

[19] When added to the model, a negative coefficient on an interaction between patience and risk-taking is significant in the baseline model but loses significance when extended controls are included (not shown).

oping composite measures of the three categories of proximate inputs. We map the input variables (see Appendix A2.2) into the three input vectors as follows: family inputs: gender, age, migration status, parental education, parental occupation, books at home, computer at home, language spoken at home, and GDP per capita (capturing overall economic wellbeing in the country); school inputs: school location, school size, share of fully certified teachers, and shortage of educational material; institutional inputs: share of privately managed schools, share of government funding at school, central exit exams, and school autonomy.

Building on the typical analysis of international educational production functions found in Woessmann (2016b), we run a pooled cross-country regression of PISA math scores on our full set of input variables. We then use the coefficient estimates on the individual variables in the model to aggregate them into family, school, and institutional factors. That is, for each input category, we calculate a linear combination as the sum of the products of the individual variables times their respective coefficients. We finally collapse the three combined input factors, as well as the residual of the achievement regression, to the country level.

This aggregation of the individual proximate input variables uses coefficient estimates from an education production function that may be biased by the omission of the deeper preference variables. Because individual coefficient estimates will be more biased for variables that are more strongly correlated with the preference measures, the estimates based on this aggregation serve as an upper bound for the preference relationships.

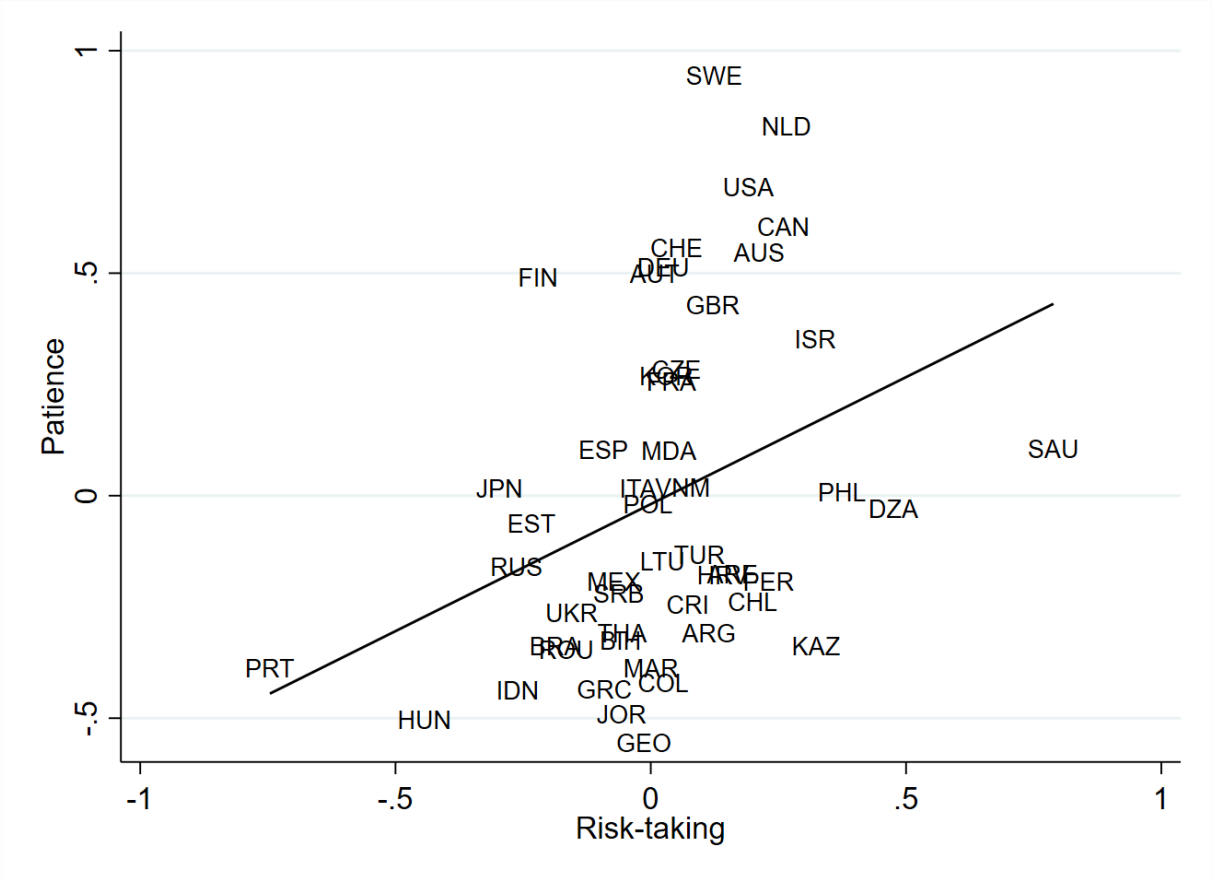
Thus, estimating the first step of the aggregation analysis including controls for the two national preferences can serve as a lower bound, as the preference measures take out important parts of the variation in the proximate inputs. By construction, the input coefficients estimated in the lower-bound analysis are unaffected by the cultural factors, and the coefficients and the R^2 for the residual category are zero.

Put together, the results of the upper-bound and lower-bound procedures presented in Table 2.4 are consistent with different input components of the education production function playing a role as channels through which the two intertemporal preferences affect student achievement.

An interesting aspect is that national preferences have limited association with institutional factors. Prior analyses have highlighted the importance of institutional factors in explaining cross-country achievement differences (Hanushek and Woessmann, 2011; Woessmann, 2016b), implying that changing institutions may be a way for nations wishing to improve their schools to break out of cultural constraints.

A2.7 Appendix Tables and Figures

Figure A2.1 : Patience and Risk-taking across Countries



Notes: Country averages. Data source: Falk et al. (2018).

Table A2.1 : Countries in the Different Analyses

			Cross-country	Migrant analysis	
	PISA	GPS	analysis	Residence country	Country of origin
	(1)	(2)	(3)	(4)	(5)
Afghanistan		x			x
Algeria	x	x	x		
Argentina	x	x	x	x	x
Australia	x	x	x	x	x
Austria	x	x	x	x	x
Bangladesh		x			x
Belarus	x			x	
Belgium	x			x	
Bolivia		x			x
Bosnia Herzegovina	x	x	x	x	x
Brazil	x	x	x		x
Brunei Darussalam	x			x	
Canada	x	x	x	x	x
Chile	x	x	x		x
China		x			x
Colombia	x	x	x		x
Costa Rica	x	x	x	x	
Croatia	x	x	x	x	x
Czech Republic	x	x	x	x	x
Denmark	x			x	
Dominican Republic	x			x	
Egypt		x			x
Estonia	x	x	x		x
Finland	x	x	x	x	x
France	x	x	x		x
Georgia	x	x	x	x	x
Germany	x	x	x	x	x
Greece	x	x	x		x
Haiti		x			x
Hong Kong	x			x	
Hungary	x	x	x		x
India		x			x
Indonesia	x	x	x	x	x
Iran		x			x
Iraq		x			x
Ireland	x			x	
Israel	x	x	x	x	
Italy	x	x	x		x
Japan	x	x	x		
Jordan	x	x	x	x	x
Kazakhstan	x	x	x		x
Kyrgyzstan	x			x	
Latvia	x			x	
Liechtenstein	x			x	
Lithuania	x	x	x		x
Luxembourg	x			x	
Macao	x			x	
Mauritius	x			x	
Mexico	x	x	x	x	
Moldova	x	x	x	x	x

(continued on next page)

Table A2.1 (continued)

	PISA	GPS	Cross-country analysis	Migrant analysis	
				Residence country	Country of origin
	(1)	(2)	(3)	(4)	(5)
Montenegro	x			x	
Morocco	x	x	x	x	x
Netherlands	x	x	x	x	x
New Zealand	x			x	
Nicaragua		x			x
Nigeria		x			x
North Macedonia	x			x	
Norway	x			x	
Pakistan		x			x
Panama	x			x	
Peru	x	x	x		
Philippines	x	x	x	x	x
Poland	x	x	x		x
Portugal	x	x	x	x	x
Qatar	x			x	
Romania	x	x	x		x
Russia	x	x	x		x
Saudi Arabia	x	x	x	x	x
Serbia	x	x	x		x
Slovakia	x			x	
Slovenia	x			x	
South Africa		x			x
South Korea	x	x	x	x	x
Spain	x	x	x		x
Suriname		x			x
Sweden	x	x	x		x
Switzerland	x	x	x	x	x
Thailand	x	x	x		x
Turkey	x	x	x	x	x
Ukraine	x	x	x	x	x
United Arab Emirates	x	x	x		x
United Kingdom	x	x	x	x	x
United States	x	x	x		x
Uruguay	x			x	
Venezuela		x			x
Vietnam	x	x	x		x
Total: 86 countries	71	64	49	48	58

Notes: The structure of country inclusion in the different parts of our analysis is complex. Three countries are included only in the baseline analysis because they participated in PISA (and GPS) but do not have migrant students with country-of-origin information (for which there is GPS data) and no student from these countries is observed as a migrant student in another PISA country. Another three countries are included in the baseline analysis and (only) as residence countries in the migrant analysis because they participated in PISA (and GPS) and have migrant students from countries of origin with GPS data, but no student from these countries is observed as a migrant student in another PISA country. 23 countries are included in the baseline analysis and both as residence countries and as countries of origin in the migrant analysis. There is also the case of 20 countries that are included in the baseline analysis and (only) as countries of origin in the migrant analysis because they participated in PISA (and GPS) but do not have migrant students with country-of-origin information (for which there is GPS data), and students from these countries are observed as migrant students in other PISA countries. 22 countries are not included in the baseline analysis, but only as residence countries in the migrant analysis because they participated in PISA, but there is no GPS data for them; however, there is GPS data for the country of origin of some of the migrant students tested in these countries. Finally, 15 countries are included only as countries of origin in the migrant analysis; these countries did not participate in PISA themselves and therefore cannot be included in the baseline analysis or as residence countries in the migrant analysis, but there is GPS data for them and students originating from these countries are observed as migrant students in residence countries that did participate in PISA.

Table A2.2 : Descriptive Statistics at the Country Level

	Mean (1)	Std. dev. (2)	Min (3)	Max (4)
PISA scores				
Math	4.520	0.560	3.524	5.410
Science	4.597	0.531	3.579	5.415
Reading	4.535	0.521	3.395	5.345
Preferences				
Patience	-0.003	0.384	-0.555	0.946
Risk-taking	0.027	0.241	-0.746	0.789
Positive reciprocity	-0.016	0.315	-1.094	0.558
Negative reciprocity	0.025	0.308	-0.510	0.716
Altruism	-0.022	0.346	-0.923	0.679
Trust	-0.016	0.249	-0.575	0.507

Notes: PISA scores: country means, pooled across all PISA waves 2000-2018, weighted by sampling probabilities. Preferences: country means of GPS preference data. Sample: 263 country-by-wave observations (reflecting 49 countries) contained in our baseline analysis of Table 2.1. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table A2.3 : Country-level Correlation of Different Preference Components

	Patience	Risk-taking	Positive reciprocity	Negative reciprocity	Altruism	Trust	WVS long-term orientation	WVS risk-taking	Hofstede long-term orientation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Risk-taking	0.358								
	(0.011)								
Positive reciprocity	-0.154	-0.148							
	(0.291)	(0.310)							
Negative reciprocity	0.236	0.334	-0.277						
	(0.103)	(0.019)	(0.054)						
Altruism	-0.051	0.110	0.699	-0.200					
	(0.728)	(0.451)	(0.000)	(0.168)					
Trust	0.197	0.162	0.259	-0.025	0.207				
	(0.176)	(0.265)	(0.072)	(0.864)	(0.153)				
WVS long-term orientation	-0.060	-0.334	-0.195	0.057	-0.163	-0.104			
	(0.700)	(0.027)	(0.204)	(0.715)	(0.290)	(0.500)			
WVS risk-taking	-0.260	0.239	0.117	0.138	0.269	0.313	-0.079		
	(0.125)	(0.160)	(0.498)	(0.423)	(0.112)	(0.063)	(0.646)		
Hofstede long-term orientation	0.247	-0.219	-0.326	0.321	-0.256	-0.246	0.609	-0.310	
	(0.115)	(0.164)	(0.035)	(0.038)	(0.101)	(0.116)	(0.000)	(0.084)	
Hofstede uncertainty avoidance	-0.558	-0.302	-0.055	0.123	-0.185	-0.527	0.006	-0.093	0.024
	(0.000)	(0.046)	(0.721)	(0.426)	(0.228)	(0.000)	(0.971)	(0.611)	(0.881)

Notes: Correlation coefficients; p-values in parentheses. Sample: 49 countries contained in our baseline analysis. Number of country observations: 49 among GPS measures, 44 between GPS and Hofstede uncertainty avoidance or WVS long-term orientation, 42 between GPS and Hofstede uncertainty avoidance and among Hofstede measures, 36 between GPS and WVS risk-taking and among WVS measures, and 32 between WVS and Hofstede measures. Data sources: Falk et al. (2018); World Values Survey (WVS); Hofstede et al. (2010).

Table A2.4 : Results by Gender

	Baseline analysis		Migrant analysis	
	Girls	Boys	Girls	Boys
	(1)	(2)	(3)	(4)
Patience	1.208***	1.242***	0.953***	0.912***
	(0.129)	(0.137)	(0.109)	(0.125)
Risk-taking	-1.190***	-1.294***	-0.302***	-0.285**
	(0.184)	(0.187)	(0.111)	(0.133)
Residence-country by wave fixed effects	No	No	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Observations	1,005,770	985,412	39,757	40,634
Countries of origin			58	57
Residence countries	49	49	48	48
R^2	0.194	0.201	0.292	0.264

Notes: Dependent variable: PISA math test score, col. 1-2: waves 2000-2018, col. 3-4: waves 2003-2018. Least squares regressions. Col. 1-2: weighted by students' sampling probability. Col. 3-4: sample: students with both parents not born in the country where the student attends school; indicated preference variables refer to country of origin. Control variables: col. 1-2: student gender, age, and migration status; imputation dummies; and wave fixed effects; col. 3-4: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; World Values Survey (WVS); Falk et al. (2018).

Table A2.5 : Model with Extended Controls

	Baseline analysis (1)	Migrant analysis (2)
Patience	0.748***	0.667***
	(0.192)	(0.100)
Risk-taking	-0.835***	-0.352***
	(0.147)	(0.092)
Residence-country by wave fixed effects	No	Yes
Baseline control variables	Yes	Yes
Extended control variables	Yes	Yes
Observations	1,992,276	80,398
Countries of origin		58
Residence countries	49	48
R^2	0.368	0.364

Notes: Dependent variable: PISA math test score, col. 1: waves 2000-2018, col. 2: waves 2003-2018. Least squares regressions. Col. 1: weighted by students' sampling probability. Col. 2: sample: students with both parents not born in the country where the student attends school; indicated preference variables refer to country of origin. Baseline control variables: col. 1: student gender, age, and migration status; imputation dummies; and wave fixed effects; col. 2: student gender, age, dummy for OECD country of origin, imputation dummies. Extended control variables, col. 1: baseline controls plus parental education, parental occupation, books at home, computer at home, language spoken at home; school location, school size, share of fully certified teachers at school, shortage of educational material; country's GDP per capita, share of privately managed schools, share of government funding at school, central exit exams, and school autonomy; col. 2: Extended control variables: baseline controls plus parental education, parental occupation, books at home, computer at home, language spoken at home; school location, school size, share of fully certified teachers at school, shortage of educational material; country-of-origin GDP per capita. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table A2.6 : Baseline Cross-country Analysis Restricted to the PISA 2015 Wave

	(1)	(2)	(3)	(4)	(5)
Patience	0.794***		1.090***	1.078***	0.763***
	(0.125)		(0.129)	(0.113)	(0.175)
Risk-taking		-0.361	-1.226***	-1.292***	-0.912***
		(0.340)	(0.220)	(0.209)	(0.178)
Positive reciprocity				0.107	
				(0.261)	
Negative reciprocity				0.289*	
				(0.158)	
Altruism				-0.235	
				(0.186)	
Trust				-0.173	
				(0.159)	
Baseline control variables	Yes	Yes	Yes	Yes	Yes
Extended control variables	No	No	No	No	Yes
Observations	319,997	319,997	319,997	319,997	319,997
Countries	41	41	41	41	41
R^2	0.102	0.013	0.157	0.171	0.329

Notes: Dependent variable: PISA math test score, wave 2015. Least squares regression weighted by students' sampling probability. Baseline control variables: student gender, age, and migration status; imputation dummies; and wave fixed effects. Extended control variables: baseline controls plus parental education, parental occupation, books at home, computer at home, language spoken at home; school location, school size, share of fully certified teachers at school, shortage of educational material; country's GDP per capita, share of privately managed schools, share of government funding at school, central exit exams, and school autonomy. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2015; Falk et al. (2018).

Table A2.7 : Results for Country Subsamples

	Baseline analysis		Migrant analysis	
	OECD	Non-OECD	OECD	Non-OECD
	(1)	(2)	(3)	(4)
Patience	0.963***	1.165**	1.028***	0.812***
	(0.180)	(0.516)	(0.105)	(0.185)
Risk-taking	-0.996***	-1.141***	-0.289**	-0.454**
	(0.271)	(0.333)	(0.132)	(0.177)
Residence-country by wave fixed effects	No	No	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Observations	1,416,506	575,770	28,519	51,879
Countries of origin			24	34
Residence countries	27	22	31	38
R^2	0.112	0.085	0.176	0.309
Difference between subsamples				
Patience	-0.202		0.216	
	(0.540)		(0.211)	
Risk-taking	0.144		0.165	
	(0.424)		(0.219)	

Notes: Dependent variable: PISA math test score, col. 1-2: waves 2000-2018, col. 3-4: waves 2003-2018. Least squares regressions. Col. 1-2: weighted by students' sampling probability. Col. 3-4: sample: students with both parents not born in the country where the student attends school; indicated preference variables refer to country of origin. Control variables: col. 1-2: student gender, age, and migration status; imputation dummies; and wave fixed effects; col. 3-4: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table A2.8 : Results in Reading and Science

	Baseline analysis		Migrant analysis	
	Science	Reading	Science	Reading
	(1)	(2)	(3)	(4)
Patience	1.121***	1.108***	0.995***	0.844***
	(0.121)	(0.113)	(0.143)	(0.144)
Risk-taking	-1.169***	-1.134***	-0.192	-0.106
	(0.180)	(0.198)	(0.124)	(0.133)
Residence-country by wave fixed effects	No	No	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Observations	1,992,276	1,950,722	80,398	80,398
Countries of origin			58	58
Residence countries	49	49	48	48
R^2	0.179	0.189	0.253	0.239

Notes: Dependent variable: PISA test score in science (col. 1 and 3) and reading (col. 2 and 4), respectively. Col. 1-2: waves 2000-2018, col. 3-4: waves 2003-2018. Least squares regressions. Col. 1-2: weighted by students' sampling probability. Col. 3-4: sample: students with both parents not born in the country where the student attends school; indicated preference variables refer to country of origin. Control variables: col. 1-2: student gender, age, and migration status; imputation dummies; and wave fixed effects; col. 3-4: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018).

Table A2.9 : Alternative WVS and Hofstede Measures of National Preferences

	Baseline analysis		Migrant analysis	
	WVS	Hofstede	WVS	Hofstede
	(1)	(2)	(3)	(4)
WVS long-term orientation	0.171*		0.176***	
	(0.091)		(0.030)	
WVS risk-taking	-0.245***		-0.120***	
	(0.075)		(0.029)	
Hofstede long-term orientation		0.339***		0.206***
		(0.054)		(0.029)
Hofstede uncertainty avoidance		-0.101		-0.092***
		(0.068)		(0.031)
Residence-country by wave fixed effects	No	No	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Observations	1,531,302	1,839,052	62,834	74,892
Countries of origin			40	48
Residence countries	36	42	44	48
R^2	0.109	0.134	0.246	0.250

Notes: Dependent variable: PISA math test score, col. 1-2: waves 2000-2018, col. 3-4: waves 2003-2018. Least squares regressions. Col. 1-2: weighted by students' sampling probability. Col. 3-4: sample: students with both parents not born in the country where the student attends school; indicated preference variables refer to country of origin. WVS and Hofstede measures z-standardized at the country level. Control variables: col. 1-2: student gender, age, and migration status; imputation dummies; and wave fixed effects; col. 3-4: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; World Values Survey (WVS); Hofstede et al. (2010).

Table A2.10 : Analysis of Unobservable Selection and Coefficient Stability following Oster (2019)

	Baseline analysis (1)	Migrant analysis (2)		
Restricted model				
Patience	1.209***	0.933***		
	(0.131)	(0.117)		
Risk-taking	-1.249***	-0.295**		
	(0.184)	(0.122)		
Observations	1,992,276	80,398		
Countries of origin		58		
Residence countries	49	48		
R^2	0.189	0.272		
Baseline model				
Patience	1.226***	0.931***		
	(0.132)	(0.116)		
Risk-taking	-1.241***	-0.294**		
	(0.184)	(0.122)		
Observations	1,992,276	80,398		
Countries of origin		58		
Residence countries	49	48		
R^2	0.198	0.275		
Oster (2019) diagnostics	Patience	Risk-taking	Patience	Risk-taking
δ to match $\beta_{1,2} = 0$	-18.093	8.224	13.575	9.761
Bound β^* for $\delta = 1$	1.358	-1.181	0.875	-0.264

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions, including 48 fixed effects for residence countries and six fixed effects for waves. Sample: students with both parents not born in the country where the student attends school. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; Falk et al. (2018).

Table A2.11 : Migrant Analysis: Models with Residence-country and Wave Fixed Effects (but not their Interaction)

	(1)	(2)	(3)
Patience (country-of-origin)	0.776***		0.929***
	(0.114)		(0.117)
Risk-taking (country-of-origin)		0.188	-0.291**
		(0.202)	(0.125)
Residence-country fixed effects	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Residence-country by wave fixed effects	No	No	No
Control variables	Yes	Yes	Yes
Observations	80,398	80,398	80,398
Countries of origin	58	58	58
Residence countries	48	48	48
R^2	0.265	0.247	0.267

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions, including 48 fixed effects for residence countries and six fixed effects for waves. Sample: students with both parents not born in the country where the student attends school. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; Falk et al. (2018).

Table A2.12 : Migrant Analysis: Subgroups by Age of Migration

	Second generation	First generation		
		All	Before age 6	After age 6
	(1)	(2)	(3)	(4)
Patience (country-of-origin)	1.023***	0.955***	1.010***	0.981***
	(0.143)	(0.120)	(0.156)	(0.103)
Risk-taking (country-of-origin)	-0.458***	-0.185	-0.228	-0.153
	(0.127)	(0.145)	(0.145)	(0.146)
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Observations	47,369	33,029	14,459	16,835
Countries of origin	56	57	51	55
Residence countries	48	48	47	48
R^2	0.297	0.263	0.298	0.258
Difference between subsamples				
Patience (country-of-origin)	-0.068		-0.029	
	(0.085)		(0.114)	
Risk-taking (country-of-origin)	0.273**		0.075	
	(0.122)		(0.062)	

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions. Sample: students with both parents not born in the country where the student attends school. Second generation: migrant students born in the country of residence. First generation: migrant students born in the country of origin; split between whether they migrated to the country of residence before or after age 6 in col. 3 and 4. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; Falk et al. (2018).

Table A2.13 : Migrant Analysis: Different Definitions of Migrants

	Narrow definition			Wide definition				Dropping countries of origin with less than 50 observations
	Mother's origin	Parental average	Separate	Mother's origin	Father's origin	Parental average	Separate	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Patience (mother's country-of-origin)	0.931***		0.343***	0.861***			0.349***	
	(0.109)		(0.069)	(0.109)			(0.070)	
Risk-taking (mother's country-of-origin)	-0.292**		0.032	-0.228*			0.038	
	(0.126)		(0.086)	(0.121)			(0.087)	
Patience (father's country-of-origin)			0.629***		0.858***		0.627***	0.939***
			(0.090)		(0.112)		(0.093)	(0.116)
Risk-taking (father's country-of-origin)			-0.339***		-0.233*		-0.336***	-0.299**
			(0.090)		(0.119)		(0.093)	(0.122)
Patience (average parents' country-of-origin)		0.941***				0.858***		
		(0.115)				(0.111)		
Risk-taking (average parents' country-of-origin)		-0.273**				-0.217*		
		(0.124)				(0.120)		
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	80,194	83,798	76,796	140,951	141,155	145,506	85,167	80,221
Countries of origin	57	58	58	60	60	60	59	46
Residence countries	48	48	48	48	48	48	48	48
R^2	0.278	0.274	0.280	0.255	0.254	0.254	0.279	0.275

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions. Sample: migrant students; see text for narrow and wide definition of migrant status. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; Falk et al. (2018).

3 Can Patience Account for Within-Country Differences in Student Achievement? A Regional Analysis of Facebook Interests*

3.1 Introduction

Human capital theory posits that activities that advance people's education can be understood as investments in skills (Becker, 1964). An important implication of this intertemporal aspect is that differences in discount rates should affect educational decisions, behaviors, and outcomes. We therefore suggest that differences in people's time preferences – patience – are an important cause of the large differences in student achievement that exist across different regions in many countries. These achievement differences are important for regional income differences; for example, skill differences account for a substantial share of income differences across U.S. states (Hanushek et al., 2017a). However, the deeper sources of this substantial regional variation in achievement are not well understood.

Investigations of whether regional differences in discount rates can account for regional variation in schooling outcomes have been stymied by a lack of region-specific measures of time preference parameters. In this paper, we exploit the massive data available from social media – in particular, Facebook interests – with machine-learning algorithms to derive new measures of regional variations in patience that permit direct assessment of the role of patience in accounting for regional differences in student achievement within countries.

Many countries have large differences in student achievement across regions. In the United States, the difference in the average math achievement of eighth-grade students on the National Assessment of Educational Progress (NAEP) between the top- and bottom-performing state is equivalent to the average learning of three school years (Hanushek et al., 2017a). A similar magnitude is found between the top- and bottom-performing region in Italy on the Istituto Nazionale per la Valutazione del Sistema Dell'Istruzione (INVALSI) test in eighth-grade math. When German states took the international test of the Programme for International Student Assessment (PISA) in 2000, state differences turned out nearly as large as international differences (Woessmann, 2010).

Since the earliest analyses of human capital, it has been recognized that discount rates constitute a fundamental determinant of individual investment decisions. But that is just part of

* This chapter is co-authored with Eric A. Hanushek, Pietro Sancassani, and Ludger Wößmann. It is based on the paper 'Can Patience Account for Within-Country Differences in Student Achievement? A Regional Analysis of Facebook Interests', mimeo.

the full impact of time preferences. Patience, the relative valuation of present versus future payoffs, appears in many decisions that relate to human capital investments. At the individual level, students must weigh current gratification such as play time with friends against study time that may lead to deferred rewards in later life. At the group level, communities and societies must trade off present against future costs and benefits when deciding how much to invest in schools, how strongly to motivate children to learn, and whether to design institutional structures to incentivize learning. Variations in patience may be relevant for understanding regional differences in educational achievement because of systematic variations of both individuals and groups across regional populations. However, the regional empirical analysis is impeded by the fact that representative survey measures of economic preferences such as patience are generally not readily available for any distinct regions within countries.

The key methodological innovation of our paper is to use social-media data to derive a measure of patience at the regional level. The underlying idea is that social-media data contain important information about people's underlying preferences such as patience. For marketing purposes, Facebook has developed an algorithm to classify the interests of over two billion people based on their observed behavior on Facebook and beyond. Specifically, self-reported interests, clicks and "likes" on Facebook, software downloads, clicks on advertisements that Facebook places on other sites, and additional inference from overall behavior and location suggest that Facebook interests post a fertile ground for investigating fundamental preferences.

Following Obradovich et al. (2022), we use dictionary vocabulary to scrape Facebook's marketing application programming interface (API) in order to derive 1,000 Facebook interests with the largest audience sizes worldwide and then use these as raw data for describing key preference differences.

Our derivation of within-country measures of patience builds on recent advances in the international analysis of culture. Expanding the approach developed by Obradovich et al. (2022), we collect data on the prevalence of Facebook interests in each country and region. After reducing the dimensionality of these by fitting a principal component analysis (PCA), we train an international model to predict the measure of patience contained in the Global Preference Survey (GPS), which developed scientifically validated measures of various preferences of country populations (Falk et al., 2018). We then use the estimated parameters of this cross-country model to predict patience for within-country regions based on their observed Facebook interests.

We validate these measures of patience by performing an international preference analysis using student achievement data from PISA. First, within the sample of GPS countries, the Facebook-derived measure performs just as well as the original GPS measure (previously used in Hanushek et al. (2022)) in predicting student achievement on the international PISA test. Second, out-of-sample prediction from the trained model allows us to expand the country sample from 48 to 80 countries. Results in the expanded sample – as well as in the subsample

of 32 new countries – are again very consistent in predicting PISA achievement. Third, both validation results are confirmed in a model that uses the subsample of migrant students and assigns them the preference parameters of their countries of origin, thereby allowing to condition on fixed effects for residence countries to shield against bias from unobserved features of students' residence countries.

We apply our method to measure patience at the regional level in two countries, Italy and the United States. In Italy, the large North-South variation across the 20 regions has raised substantial interest in policy and research (e.g., Putnam et al. (1992); Ichino and Maggi (2000); Guiso et al. (2004)). As a large federal country, the United States allows for regional analyses for a large sample of 50 U.S. states. Both countries show substantial regional variation in the Facebook-derived measure of patience with a noteworthy North-South gradient.

We employ the newly derived regional measure of patience in analyses of regional student achievement in the two countries. For Italian regions, we use achievement data from over 200,000 students on the national INVALSI test. For the United States, we use regional achievement data on the national NAEP test. By studying achievement differences for regions within individual countries, the estimation is less prone to confounding from unobserved national traits such as languages, constitutions, and institutional factors that may hamper prior cross-country analyses.

In both countries, regional differences in patience account for substantial parts of the sub-national variation in student achievement. The models account for over two thirds of the variation in test scores across Italian regions and for over one third across U.S. states. The smaller role in the United States may reflect that the substantial internal mobility of the U.S. population across states might introduce attenuation bias in the regional measurement of intergenerationally transmitted cultural traits.

Consistent with skill development as a cumulative process, the estimated association of patience with student achievement increases across grade levels. In the Italian INVALSI tests, estimates grow steadily across the four testing occasions from second to tenth grade. Similarly, estimates for the U.S. NAEP are smaller in fourth than in eighth grade.

Results are stable in a series of robustness analyses such as using reading achievement or the regionally representative participation of Italy in PISA 2012. Throughout, our analysis conditions on regional variation in risk-taking, another preference parameter that can partly capture intertemporal aspects. However, the machine-learning model to predict risk-taking from Facebook interests does not perform very well at the regional level. As patience and risk-taking tend to be positively associated and prior work suggests a negative association of risk-taking with student achievement, the poor measurement of risk-taking may imply that the estimates of patience reflect lower bounds.

Our analysis contributes to two strands of literature. First, we contribute to the analysis of the role of time preferences in human capital investment. Our regional analysis adds a new perspective to the literature that has studied the role of patience for educational outcomes at the individual level (Sutter et al., 2013; Golsteyn et al., 2014; Castillo et al., 2019) and at the international level (Figlio et al., 2019; Hanushek et al., 2022). Additionally, cross-country work has shown the importance of patience for long-run comparative economic development (Galor and Özak, 2016; Sunde et al., 2022). In deriving the regional patience measure, our approach also contributes to the literature that uses Facebook data to measure various concepts of culture and social networks (e.g., Obradovich et al. (2022); Chetty et al. (2022); Bailey et al. (2022)), as well as to the literature on culture and economic outcomes more broadly (e.g., Guiso et al. (2006); Alesina and Giuliano (2015)). Second, the consideration of patience contributes a new perspective of deeper causes to the study of regional differences in student achievement. While there are a few studies on proximate causes such as family background, school spending, and institutional settings (e.g., Hanushek and Raymond (2005); Woessmann (2010); Dee and Jacob (2011)), most stop at just noting the magnitudes of regional differences without providing convincing explanations of them (e.g., Hanushek (2016)).

The remainder of the paper is structured as follows. Section 3.2 describes our method to derive regional measures of patience from data on Facebook interests and includes a validation exercise at the cross-country level. Section 3.3 describes the regional student achievement data. Section 3.4 presents our results. Section 3.5 concludes.

3.2 Methods: Deriving Regional Patience Measures from Facebook Interests

We use social-media data to measure patience at the regional level. Section 3.2.1 introduces the Facebook interest data. Section 3.2.2 validates the suitability of these interests to predict international differences in patience. Section 3.2.3 describes our method to derive regional measures of patience from the Facebook interests.

3.2.1 Facebook Interests

With 2.9 billion monthly active users, Facebook is the world's largest social network.[1] Facebook's core business consists of selling advertising space on its social media platform. In 2021, 97.5 percent of Facebook's revenues came from advertisements.[2] Hence, Facebook's business model depends primarily on its ability to keep users engaged on the platform while

[1] Source: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (last accessed 23 February 2023).

[2] Figures about Facebook's users and revenues are reported by Meta, Facebook's parent company, drawing on the third-quarter 2022 results (https://s21.q4cdn.com/399680738/files/doc_financials/2022/q3/Meta-09.30.2022-Exhibit-99.1-FINAL.pdf, last accessed 2 January

advertisers promote their products and services to users who may find them relevant. To this purpose, Facebook puts considerable effort into inferring users' interests (Thorson et al., 2021).

Facebook determines users' interests using a variety of sources, both inside the Facebook platform as well as on external websites (Cabañas et al., 2018; Obradovich et al., 2022). Inside the Facebook platform, these sources include personal information that users share on Facebook as well as users' activity on Facebook, such as page likes, group membership, and content users engage with. Outside the platform, Facebook tracks users' visited websites, installed apps, and purchasing behavior.[3] Facebook uses these data to deliver content and recommendations based on users' interests and to allow advertisers to target users whose interests are relevant for the products or services that they want to sell.[4]

The hundreds of thousands of interests classified by Facebook are organized in nine main categories: business and industry, entertainment, family and relationships, fitness and wellness, food and drink, hobbies and activities, shopping and fashion, sports and outdoors, and technology. Interests can be very broad, such as 'Entertainment' or 'Music', or very narrow, such as 'Caribbean Stud Poker', a casino table game. Figure 3.1 shows the 1,000 Facebook interests with the largest worldwide audience, where larger font sizes correspond to larger audience sizes. Interests often relate to leisure activities such as sports and beauty, but also to broader categories such as education and politics.

Following Obradovich et al. (2022), we proceed in two steps to retrieve data on the Facebook interests for countries and subnational entities. First, we obtain a comprehensive list of Facebook interests by querying the Facebook Marketing API, the interface that allows advertisers to configure their advertisement campaigns. For any given text input (query), a tool within the API returns a collection of the respective closely related Facebook interests together with their estimated worldwide audience and a unique identifier, which makes them language-independent. We iteratively feed this function with all 25,322 terms of an English dictionary[5] and 2,000 randomly selected titles of Wikipedia articles, each of which can yield

2023) and the 2021 annual report (https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/14039b47-2e2f-4054-9dc5-71bcc7cf01ce.pdf, page 58, last accessed 2 January 2023).

[3] While official figures on Facebook's off-platform data collection are not available, Aguiar et al. (2022) estimate that, for a representative sample of 5,000 U.S. internet users in 2016, Facebook can track 55 percent of websites visited by Facebook users, which amounts to 41 percent of browsing time. For more information on this practice, see also Facebook's official press release on data collection outside of Facebook at https://about.fb.com/news/2018/04/data-off-facebook/ (last accessed 2 January 2023).

[4] Facebook users can access the interests that Facebook assigns to them. According to a recent report, 59 percent of Facebook users in the US say that these Facebook interests reflect their real-life interests (https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/, last accessed 23 February 2023).

[5] We use a dictionary of popular English words available at https://github.com/dolph/dictionary/blob/master/popular.txt (last accessed 3 January 2023).

several Facebook interests. After removing duplicates, we obtain a collection of 41,513 unique interests from this procedure.

Second, we select the 1,000 interests with the largest worldwide audience obtained in the previous step, which ensures cross-country and within-country comparability. For each of these 1,000 interests, we again use the tool from Facebook's Marketing API to separately obtain the estimated audience size for each country in which Facebook has a presence, as well as for each state in the U.S. and region in Italy. For each geographical entity, this process yields a vector of size 1,000 with the estimated audience for all of the 1,000 largest interests by worldwide audience. Finally, we divide the estimated audience by the 2020 population size in each geographical entity to obtain the share of individuals holding each interest.

3.2.2 Using Facebook Interests to Measure Patience: A Cross-country Validation Exercise

To assess the suitability of the Facebook interest data to measure patience, we perform a cross-country validation exercise which proceeds in four steps. First, we reduce the dimensionality of the Facebook data. Second, we study how well the reduced-dimensionality Facebook data predicts an external measure of patience available at the country level in the Global Preference Survey (GPS). Third, after training the prediction model within the sample of GPS countries, we perform out-of-sample predictions to expand the country sample to countries that are not part of the GPS. Fourth, we use the international PISA test data to validate whether the Facebook-derived measure of patience is associated with student achievement across countries both within and outside the sample of countries participating in the GPS.

We start by reducing the dimensionality of the country-level Facebook interests by a principal component analysis (PCA) fitted on the international sample of all 216 countries and geographical entities featured by Facebook. On top of reducing the dimensionality of the variables that we use to later train the machine-learning model, this step also avoids collinearity problems because the resulting principal components are uncorrelated by construction. The first 10 principal components (PCs) capture 70 percent of the total cross-country variance contained in the Facebook interests, the first 20 PCs capture 80 percent, and the first 48 PCs capture 90 percent.[6] While the additional variance captured by any PC beyond the 10th PC is quite small, this still suggests that many PCs are required to capture the full variance in Facebook interests across countries (see also Obradovich et al. (2022)).

Next, we train a machine-learning model to learn the relationship between the country-level PCs of the Facebook interests and an external measure of the countries' patience. As an external measure, we use the measure of patience contained in the GPS, which collected survey-based measures of several preference parameters from representative samples in 76 countries (Falk et al., 2018). The measure of patience combines a qualitative survey item

[6] Details are provided in Appendix Figure A3.1.

and a hypothetical choice scenario that were chosen based on their predictive capacity for incentivized choices in an ex-ante laboratory setting. Our training sample includes 74 countries, namely all the countries that participated in the GPS survey except for Iran and Russia, for which Facebook data are currently not available.[7] We use a 10-fold cross-validated least absolute shrinkage and selection operator (LASSO) model for the cross-country training. The performance of the model is quite satisfactory: Independent of whether 10, 20, 30, 40, 50, or even 100 PCs are used, the R^2 of the in-sample prediction of patience by the reduced-dimensionality Facebook interests is quite stable between 0.65 and 0.70.[8]

We use the parameter estimates of the machine-learning model to make out-of-sample predictions of patience for all countries that participated in at least one PISA wave and for which Facebook interests can be retrieved. Given the limited size of the sample used to train the machine-learning model, we prefer the most parsimonious specification with 10 PCs for the out-of-sample predictions to avoid overfitting.[9] The resulting sample for which we have both Facebook-derived patience measures as well as student test scores consists of 80 countries.[10]

The R^2 of the in-sample prediction for risk-taking is somewhat lower than for patience,[11] which suggests that risk-taking is harder to predict from Facebook interests compared to patience in the cross-country setting.

To validate our Facebook-derived measures of patience and risk-taking, we estimate their relationship with student achievement across countries. The model setup for the validation follows Hanushek et al. (2022), using math achievement on the PISA test over all seven available waves 2000-2018 to estimate the following OLS model:

$$T_{ict} = \beta_1 Patience_c + \beta_2 Risk_c + \alpha_1 \mathbf{B}_{ict} + \mu_t + \varepsilon_{ict} \tag{3.1}$$

where T, the standardized PISA test score of student i in country c in year t, is a function of the country-level measures of patience and risk-taking of country c, a vector of control variables B (student gender, age, and migration status), and an error term ε_{ict}. Fixed effects for test waves μ_t account for time trends and idiosyncrasies of the individual tests. The coefficients of interest are β_1 and β_2 which characterize the relationship of patience and risk-taking with student achievement. Regressions are weighted by students' sampling probability, giving equal weight to each country. Standard errors are clustered at the country level.

[7] A list of the countries is shown in column 4 of Appendix Table A3.1.

[8] Details are provided in Appendix Figure A3.2.

[9] Less parsimonious models tend to obtain better in-sample performance (although this is hardly the case for patience, see Appendix Figure A3.2) but can lead to worse out-of-sample performance especially with small samples.

[10] The countries are reported in columns 1 and 3 of Appendix Table A3.1.

[11] See Appendix Figure A3.2.

The Facebook-derived measures of patience and risk-taking perform very well in the cross-country validation exercise. As a baseline, the first column of Panel A of Table 3.1 shows that patience has a strong and significant positive relationship with student achievement when using the original GPS measure, whereas risk-taking has a strong and significant negative relationship.[12] Column 2 substitutes the GPS measures of patience and risk-taking with our Facebook-derived measures, using the same sample of countries.[13] The results are very much in line with those obtained using the original GPS measures, which corroborates the validity of the Facebook-derived measures. Point estimates are in fact slightly larger (in absolute terms) than the original estimates.[14] The out-of-sample predictions allow us to extend the analysis of the Facebook-derived measures of patience and risk-taking from a sample of 48 to 80 countries – all countries that participated in PISA and have Facebook data – encompassing over 2.6 million student observations. Results generalize very well to the extended sample, with increased precision and without significantly different estimates (column 3). Even in the sample of 32 countries that were not part of the original GPS analysis, results are qualitatively the same and statistically highly significant (column 4).

In the international analysis, we can also perform a migrant analysis that aims to account for unobserved differences across residence countries. The analysis restricts the sample to students with a migrant background and assigns them the values of patience and risk-taking of their home countries (see Figlio et al. (2019); Hanushek et al. (2022)). By observing migrant students from different countries of origin who are schooled in the same residence country, this setup allows to take out fixed effects of the residence countries (as well as their full interaction with wave fixed effects), thereby excluding the possibility that the relationships are driven by other factors of the country of schooling.

The migrant analysis further validates the informational content of the Facebook-derived measures. Results in Panel B of Table 3.1 show that again, the positive patience relationship and the negative risk-taking relationship again replicate very well when using the Facebook-derived rather than the original GPS measures.[15] The risk-taking coefficient is somewhat less precisely estimated but actually increases in (absolute) size. Estimates become quite imprecise (and larger) when restricting the sample to non-GPS countries (column 4), indicating limited power of the migrant analysis in the smaller sample.

[12] This model replicates the main estimates of Hanushek et al. (2022) after dropping Russia (which has no Facebook data), with estimates hardly changed (see column 3 of their Table 1).

[13] The measures are obtained with 10 PCs of Facebook interest. Appendix A3.1 shows that results are very similar when using additional (20-50) PCs to derive the measures.

[14] The coefficient on patience in column 2 of Table 3.1 is significantly larger than in column 1 in the cross-country analysis, whereas all other differences between columns 1 and 2 are statistically insignificant.

[15] With the Facebook data, we expand the countries of origin considered in the migrant analysis from 56 to 93 (see Appendix Table A3.2). The destination countries increase only from 46 to 50 because other PISA countries do not report students' and parents' country of birth required to determine migrants' country-of-origin preferences.

Overall, the cross-country validation exercise shows that the measures of patience and risk-taking predicted using the Facebook data follow very closely the patterns from externally validated survey measures of these preferences. This implies that the information contained in the Facebook interests and their underlying principal components are suitable to infer such measures for geographical entities that do not have representative measures from surveys.

3.2.3 Predicting Regional Patience from Reduced-Dimensionality Facebook Interests

Our method to derive measures of patience for subnational regions from the Facebook interests, which extends the method developed by Obradovich et al. (2022) to our regional analysis, proceeds in three steps. First, we again reduce the dimensionality of the Facebook interests using a PCA, but this time fitting the PCA across the regions within a given country. Second, we use the PC loadings obtained from the within-country PCA to reduce the dimensionality of the country-level Facebook interests in the international sample. This allows us to train a machine-learning model that learns the relationship between these country-level PCs and the survey-based measure of patience contained in the GPS. Third, we use the parameter estimates from the internationally trained machine-learning model with the PC loadings derived from fitting the PCA at the regional level to make out-of-sample predictions of patience for the subnational regions based on their Facebook interests.

We fit the PCA to reduce the dimensionality of the Facebook interests separately within the two countries we study, i.e., for Italian regions and for U.S. states. Fitting the PCA at the regional level ensures that the PCs capture variance in dimensions of Facebook interests that are relevant at the regional level within the specific country. For the Italian regions, the first 10 PCs already capture 90 percent of the within-country variance in Facebook interests.[16] For the U.S. states, the same portion of variance is captured by the first 15 PCs. In both cases, each subsequent PC only captures a small portion of variance.

To train a prediction model of the country-level patience measures, we first apply the respective within-country PCA to the international sample. That is, we use the PC loadings obtained in the previous step for dimensionality reduction of the country-level Facebook interests. Because these PC loadings capture the contribution of the regional-level Facebook interests to the PCs, the resulting country-level PCs will preserve the respective variance that can be found in Facebook interests across Italian regions or U.S. states. We then use these PCs to train a 10-fold cross-validated LASSO model to learn the relationship between the PCs and the GPS measure of patience across countries.[17] Since the country-level PCs are now constructed to resemble the regional-level variance in Facebook interests, the model should be capable

[16] Details are provided in Appendix Figures A3.3 and A3.4 for Italy and the United States, respectively.

[17] The GPS measure is standardized to have mean zero and standard deviation one across individuals in the 76 countries participating in the GPS, so that estimates in our subsequent analysis can be interpreted in terms of standard deviations.

of generalizing the estimated relationship between country-level PCs and countries' GPS measures to Italian regions and U.S. states.

The in-sample performance of the model in predicting the GPS measure of patience is relatively good, both when the PC loadings are derived from fitting the PCA on the Facebook interests of Italian regions and of U.S. states. Few PCs already capture a considerable portion of the variation in Facebook interests within countries: with 10 PCs, the R^2 of the in-sample prediction reaches 0.5 in the case of Italian regions and over 0.6 in the case of U.S. states.[18] In both cases, increasing the number of PCs and, hence, the amount of variance used, is accompanied by an increase in the in-sample performance of the model, but we again prefer more parsimonious models for the out-of-sample predictions to avoid overfitting.

We then derive regional measures of patience by using the parameter estimates from the internationally trained model to predict patience from the Facebook interests observed in Italian regions and U.S. states, respectively. Figure 3.2 contains maps that show the regional variation of the Facebook-derived measure of patience in Italy and the United States.[19] In Italy, the regions in the lowest deciles of patience are Sicily and Campania in the South. The region with the highest level of patience is Trentino-Alto-Adige, located in the North-East. Interestingly, parts of Trentino-Alto-Adige belonged to Austria and the former Austro-Hungarian empire for long periods of time, and large parts of the population in the region speak German as their first language. According to the country-level GPS measures, Austria has a much higher level of patience than Italy.[20] The fact that this region exhibits the largest level of patience thus bodes well for the Facebook-derived measure. In the United States, the states that exhibit the highest level of patience are Vermont and Maine in the North-East. Both countries tend to show a North-South gradient in the Facebook-derived measure of patience.

When performing the same prediction analysis for risk-taking, the performance of the prediction model is substantially worse. Both for Italian regions and for U.S. states, the R^2 of the in-sample prediction is well below 0.2 for all models with up to 10 PCs and well below 0.4 even for a model with 20 PCs.[21] We include the measure of risk-taking as a control variable in our regional analysis throughout.[22] However, its poor measurement when PC loadings are fitted at the regional level means that the estimates on patience are likely lower bounds because

[18] Details are provided in Appendix Figures A3.5 and A3.6 for Italy and the United States, respectively.

[19] The figure shows values obtained with 4 PCs; patience measures obtained with different numbers of PCs yield the same graphical representation.

[20] The country-level GPS measure of patience for Austria (0.61) is half a standard deviation higher than for Italy (0.11). A similar argument can be made for the Aosta Valley region in the North-West of Italy, whose culture is deeply intertwined with neighboring France. France's GPS measure of patience is a quarter of a standard deviation higher than Italy's.

[21] See Appendix Figures A3.5 and A3.6. The performance with 20 PCs is a spike that likely reflects overfitting of the data in this case.

[22] See Appendix Figure A3.7 for maps depicting the regional distributions of risk-taking in Italy and the United States, but these should be interpreted with care because of the poor performance of the prediction model.

patience and risk-taking are positively associated and risk-taking has the opposite sign from patience in the cross-country analysis (Hanushek et al., 2022).

3.3 Data on Regional Student Achievement

To estimate the association of patience with student achievement for subnational regions, we use data on the largest student assessments for Italy and the United States, respectively, that are both representative at the regional level: INVALSI (Section 3.3.1) and NAEP (Section 3.3.2).

3.3.1 Italy: INVALSI

Since 2007, the Istituto Nazionale per la Valutazione del Sistema Dell'Istruzione (INVALSI) assesses a random sample of Italian students in math and Italian every year. Furthermore, INVALSI administers student, teacher, and principal questionnaires to collect background information about the educational environment. We use data on math achievement in the school years 2017-2018 and 2018-2019, the last years before the COVID-19 pandemic. In our main analysis, we focus on eighth-grade students since they are closest in age to the students in PISA and NAEP. The sample of eighth-graders consists of 59,034 students. In additional analyses, we also use data for students in grades 2, 5, and 10, with an entire sample size of 235,661 students.

The random sample of students is drawn following a two-step procedure, where a varying number of classes is randomly selected within a random sample of schools stratified at the regional level. Crucially for our analysis, the sample is representative at the regional level for 19 of the 20 regions in Italy (Falorsi et al., 2019). The exception is Trentino-Alto-Adige, where only students in the autonomous municipalities of Bolzano and Trento are tested. The difference between the lowest and highest performing region in Italy in 8th-grade math amounts to roughly three quarters of a standard deviation, equivalent to the average learning of almost three school years.

In robustness checks, we complement the INVALSI analysis using Italian data from PISA 2012 where Italy oversampled students in each region to obtain a representative sample of students.

3.3.2 United States: NAEP

We use data from the National Assessment of Educational Progress (NAEP), the largest nationally representative assessment of students in the United States. In our main analysis, we focus on NAEP mathematics test scores in grade eight. We combine mathematics test scores into a single average score for each state using data from the last three waves of NAEP before the COVID-19 pandemic, namely NAEP 2015, 2017 and 2019. The resulting dataset consists of state-level test scores for the 50 U.S. states and the federal district of Washington,

D.C. Approximately 140,000 students take part in a typical NAEP assessment.[23] In additional analyses, we also use data on fourth-grade students. Also in the United States, the difference between the lowest and highest performing state in 8th-grade math is equivalent to roughly three years of schooling.

We divide both INVALSI and NAEP test scores by the student-level standard deviation in the respective country, so that regression coefficients can be interpreted in terms of standard deviations.

3.4 Results

We use our regional measure of patience derived from Facebook interests to study whether differences in patience can account for the substantial differences in student achievement that exist across Italian regions and U.S. states. The estimated models are versions of equation (3.1) applied to the regional rather than the country level.[24] Compared to the cross-country analysis, the within-country analysis is less prone to bias that may arise from national factors such as languages, laws, and institutional settings. In this section, we report our results for Italy (Section 3.4.1) and the United States (Section 3.4.2), followed by robustness analyses (Section 3.4.3).

3.4.1 Italy

Italy represents an interesting case study for the regional analysis because of its well-known North-South divide in many dimensions, including student test scores. This regional divide is surprising given the relatively centralized structure of the country: the schooling system is regulated mostly at the country level, having the same structure across regions.[25] Hence, the within-country association between patience and student test scores is unlikely to be severely biased by institutional factors.

The Facebook-derived regional measure of patience is strongly and significantly associated with student achievement across Italian regions. Panel A of Table 3.2 shows results of student-level analyses of math achievement in eighth grade using patience measures obtained with 4, 7, and 10 PCs of Facebook interest, which showed good in-sample performance in Section 3.2.3. Irrespective of the number of PCs used to derive the patience measure, the coefficient

[23] Source: `https://nces.ed.gov/nationsreportcard/guides/statsig.aspx` (last accessed 23 February 2023).

[24] The model specification is very parsimonious as we think of patience as a deep determinant of student achievement. Proximate inputs often included in education production functions such as parental education or school resources would be bad controls in this setting as they are endogenous to regions' patience.

[25] The matters in which the state has exclusive legislation are listed in Article 117 of the Italian Constitution (`https://www.governo.it/it/costituzione-italiana/parte-seconda-ordinamento-della-repubblica/titolo-v-le-regionile-province-e-i`; last accessed 30 January 2023).

estimates are highly significant and indicate that a one standard deviation (SD) increase in patience is associated with an increase in math test scores by 1.35-1.51 SD. The Italian regional estimates are only slightly smaller than the cross-country estimates reported in Table 3.1.

When estimated at the regional level, results suggest that regional differences in patience can account for at least two thirds of the variation in student achievement across Italian regions. Using student test scores aggregated to the regional level in Panel B of Table 3.2, point estimates are very similar, albeit slightly smaller than in the student-level analysis. The R^2 indicates that the model accounts for 0.68-0.80 of the region-level variation, indicating that patience accounts for a large portion of the differences in student achievement across Italian regions.

Interestingly, the association of patience with student achievement increases strongly with increasing grade levels. Panels A and B of Table 3.3 show results for all four grade levels available in INVALSI for the patience measure obtained with 4 PCs of Facebook interests.[26] Column 3 replicates our main results from the previous table that refer to students in grade 8. The other columns show results for students in grades 2, 5, and 10, respectively. Coefficient estimates increase continuously from an insignificant 0.29 SD in grade 2 to a highly significant 1.77 in grade 10 when estimated at the student level. Region-level estimates are again quite similar. These results suggest that as educational investments are cumulative, the role of patience keeps adding up across grades.

3.4.2 United States

As a large federal country, the United States provide a large regional sample of 50 states plus Washington, D.C that feature large differences in student outcomes.[27] With data accessible only at the state level, Panel C of Table 3.2 reports the results of our state-level regressions. The analysis again refers to math achievement in 8th grade and uses Facebook-derived measures of patience obtained with 4,7, and 10 PCs.

Also in the United States, patience is significantly associated with higher student achievement at the regional level. A one SD increase in the Facebook-derived measure of patience is associated with an increase of 0.17-0.29 SD in test scores across U.S. states. The point estimates are only about a quarter of the ones estimated for Italian regions. The model accounts for slightly more than one third of the variation in test scores across U.S. states.

While patience plays an important role in accounting for cross-state differences in student test scores in the United States, the role is less prominent than in Italy. A possible explanation is that the population in the United States is substantially more mobile and mixed. According to census estimates, 42 percent of the U.S. population lives in a state different from their state

[26] Results are very similar when using 7 or 10 PCs (not shown).
[27] Results are similar when excluding Washington, D.C. from the analysis (not shown).

of birth.[28] Because cultural traits such as patience are mostly transmitted across generations (e.g., Bisin and Verdier (2011); Alesina and Giuliano (2014)), such an extent of internal migration makes cultural traits harder to measure at the state level. This might induce measurement error in the estimates of patience and cause attenuation bias in the regressions.

Consistent with the Italian evidence, the association between patience and student achievement is smaller in lower grades also in the United States. While also statistically significant, the coefficient estimate in 4th grade is only about half the size as in 8th grade (Panel C of Table 3.3), corroborating that the role of patience adds up as educational efforts accumulate.

3.4.3 Robustness Analysis

Results prove stable in a series of robustness analyses. Both in Italy and the United States, we find similar results for reading achievement, with slightly smaller point estimates. Results are also robust in the separate waves available in both countries. They show similarly for girls and boys, with no significant gender difference.

The availability of individual-level data for Italy allows for additional in-depth analyses. Consistent with a leading role of culture, estimates are larger for native students than for migrant students. Results are robust to excluding Trentino-Alto-Adige whose sample is not representative for the entire region and whose German-language population might limit comparability. Results are also robust in an Oster (2019) analysis of unobservable selection and coefficient stability. Furthermore, results are remarkably similar when using Italian regional performance on the PISA 2012 test. Appendix A3.1 provides the details of these robustness analyses, together with the respective estimation tables.

3.5 Conclusions

Regional differences in student achievement are poorly understood and understudied. In this paper, we deploy social-media-derived measures of time preferences to provide evidence that patience can account for large portions of such differences. We first show that our Facebook-derived measures perform just as well as scientifically validated survey measures of patience and risk-taking when studying cross-country differences in student achievement. We leverage the broader coverage of our new measures to show that patience and risk-taking are strongly associated with student test scores in a much larger sample of countries than previously studied.

In our regional analysis of Italy and the United States, we test the extent to which patience can account for differences in student achievement across regions. We find that even within

[28] Own calculations based on the ACS 2019 table of state of residence by place of birth available at `https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-of-residence-place-of-birth-acs.html` (last accessed 25 February 2023).

countries, where schooling systems and educational inputs tend to be more homogenous than between countries, patience is strongly positively associated with student test scores. The model can account for over two thirds of the regional variation in student achievement in Italy and over one third in the United States.

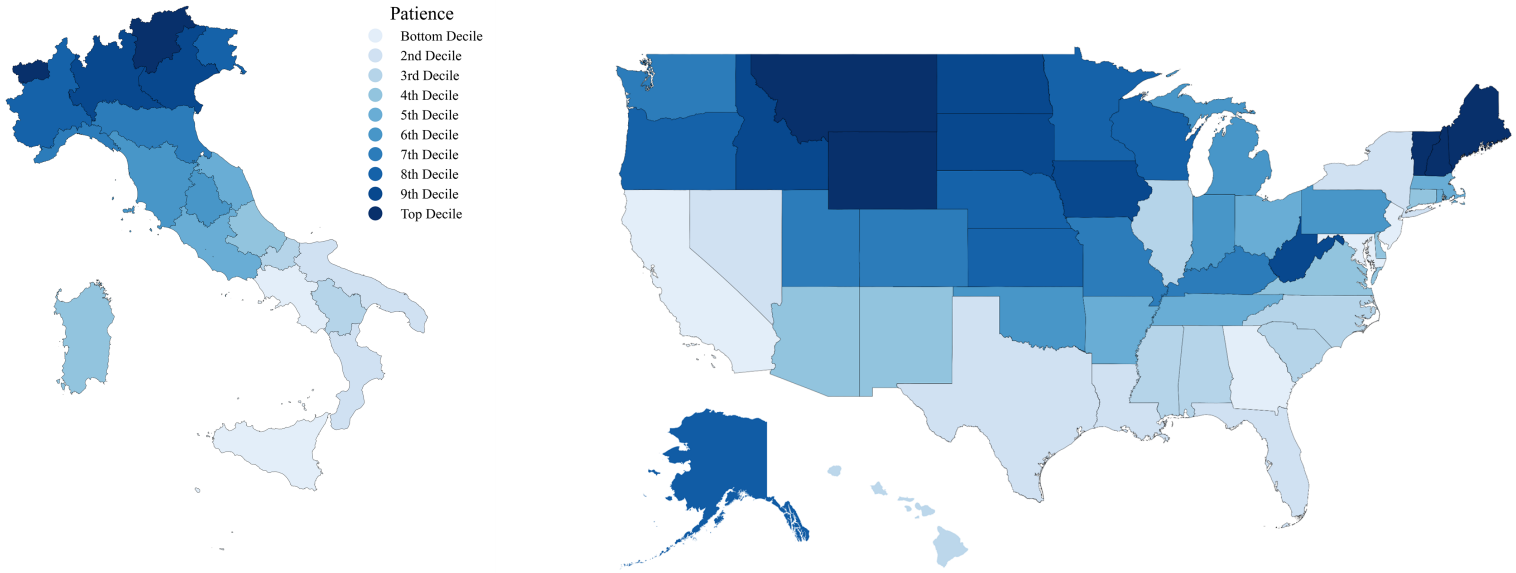
Our findings imply that due to differences in patience, similar educational inputs can lead to substantially different outcomes. When addressing within-country differences in student achievement, policymakers should therefore take possible differences in patience into account. While cultural traits are considered hard to change (e.g., Guiso et al. (2006); Bisin and Verdier (2011)), recent evidence shows that traits such as patience are malleable, especially at a young age, and can be improved through specific interventions (e.g., Bird (2001); Alan and Ertac (2018); Jung et al. (2021)). Hence, policies aimed at increasing patience seem a promising avenue to address regional deficits in student outcomes.

Figures and Tables

Figure 3.1 : Word Cloud of Facebook Interests

Notes: Word cloud of 1,000 Facebook interests with largest worldwide audience. Larger font sizes indicate larger audience.

Figure 3.2 : Patience, Risk-taking, and Student Achievement across Countries.



A. Italian Regions **B. U.S. States**

Notes: The figures show maps of the Facebook-derived measure of patience obtained with 4 PCs for Italian regions (Panel A) and U.S. states (Panel B), respectively. Each color corresponds to a decile of the distribution of patience within each country. Darker colors denote higher levels of patience.

Table 3.1 : Patience, Risk-taking, and Student Achievement: Cross-Country Validation Exercise

	GPS measure	Facebook measure (10 PCs)		
		Original sample	Extended sample	Non-GPS sample
	(1)	(2)	(3)	(4)
A. Cross-country analysis				
Patience	1.225***	1.673***	1.712***	1.761***
	(0.132)	(0.134)	(0.118)	(0.209)
Risk-taking	-1.229***	-1.336***	-1.507***	-1.625***
	(0.188)	(0.304)	(0.249)	(0.378)
Control variables	Yes	Yes	Yes	Yes
Observations	1,954,840	1,954,840	2,660,408	705,568
Residence countries	48	48	80	32
R^2	0.200	0.210	0.220	0.241
B. Migrant analysis				
Patience	0.957***	0.805***	0.902***	1.766***
	(0.115)	(0.182)	(0.205)	(0.481)
Risk-taking	-0.315**	-0.677**	-1.221***	-3.531***
	(0.124)	(0.278)	(0.350)	(0.549)
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	78,403	78,403	90,983	12,580
Countries of origin	56	56	93	37
Residence countries	46	46	50	34
R^2	0.280	0.272	0.298	0.310

Notes: Dependent variable: PISA math test score. Least squares regressions. Panel A: all PISA waves 2000-2018; weighted by students' sampling probability. Panel B: waves 2003-2018; students with both parents not born in the country where the student attends school; including 180 fixed effects for each residence-country by wave cell. Control variables: Panel A: student gender, age, and migration status; imputation dummies; and wave fixed effects; Panel B: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level (migrant analysis: country of origin) in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018); own elaboration of Facebook data.

Table 3.2 : Patience and Student Achievement: Regional Analysis for Italy and the United States

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
A. Italy (individual level)			
Patience	1.505***	1.350***	1.437***
	(0.197)	(0.114)	(0.117)
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	59,034	59,034	59,034
Regions	20	20	20
R^2	0.092	0.099	0.099
B. Italy (regional level)			
Patience	1.246***	1.134***	1.207***
	(0.193)	(0.095)	(0.099)
Wave fixed effects	Yes	Yes	Yes
Observations	42	42	42
Regions	20	20	20
R^2	0.679	0.790	0.795
C. United States (state level)			
Patience	0.293***	0.172*	0.285**
	(0.089)	(0.096)	(0.132)
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
Regions	51	51	51
R^2	0.360	0.348	0.360

Notes: Dependent variable: Panels A and B: INVALSI 8th-grade math test score in waves 2018 and 2019; Panel C: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination; Panel C: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional (state) level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table 3.3 : Patience and Student Achievement at Different Grade Levels

	Grade 2	Grade 4/5	Grade 8	Grade 10
	(1)	(2)	(3)	(4)
A. Italy (individual level)				
Patience	0.291	0.534*	1.505***	1.767***
	(0.193)	(0.286)	(0.197)	(0.236)
Control variables	Yes	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes	Yes
Observations	48,812	50,608	59,034	77,207
Regions	20	20	20	20
R^2	0.028	0.032	0.092	0.151
B. Italy (regional level)				
Patience	0.182	0.365	1.246***	1.466***
	(0.202)	(0.237)	(0.193)	(0.247)
Wave fixed effects	Yes	Yes	Yes	Yes
Observations	42	42	42	42
Regions	20	20	20	20
R^2	0.044	0.075	0.680	0.678
C. United States (state level)				
Patience	-	0.156**	0.293***	-
		(0.064)	(0.089)	
Wave fixed effects		Yes	Yes	
Observations		153	153	
Regions		51	51	
R^2		0.158	0.360	

Notes: Dependent variable: Panels A and B: INVALSI 8th-grade math test score in waves 2018 and 2019; Panel C: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination; Panel C: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional (state) level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Appendix

A3.1 Robustness Analysis

This appendix reports a series of robustness checks for the cross-country validation exercise (Appendix 3.1.1), for the analysis of Italian regions (Appendix 3.1.2), and for the analysis of the U.S. states (Appendix 3.1.3). The analysis of the cross country-validation exercise shows that results do not depend on the specific procedure used to derive the measures of patience and risk-taking. For Italy and the United States, the analysis shows that results are robust to different student outcomes and across various subsamples. The availability of individual-level data for Italy allows a more in-depth analysis than for the United States, where the analysis is constrained by the regional-level data.

Cross-Country Validation Exercise

To make sure that the results of the validation exercise in Section 3.2.2 do not depend on the specific way of predicting patience and risk-taking from the Facebook data, we present results for alternative predictions that vary the number of PCs used in the LASSO that predict patience and risk-taking from the Facebook interests. Table 3.1 in the main text shows results using the first 10 PCs resulting from the PCA performed on the international sample of Facebook interests. Here, we report variations of up to the first 50 PCs.

Table A3.3 shows the results from alternative predictions of patience and risk-taking for the cross-country analysis. Columns 1-4 report results when using the first 20, 30, 40, and 50 PCs, respectively, when predicting the two traits in the international sample. Panel A performs the analyses for the sample of 48 countries that participated in the GPS. Panel B shows the same analyses for the extended sample of 80 countries. Results are qualitatively and quantitatively very similar to the respective results in Table 3.1, which implies that the relationship between the Facebook interests and the two cultural traits is very stable in the international sample.

Table A3.4 shows the equivalent results for the same variation in PCs in the migrant analysis. The results for patience are stable across the different numbers of PCs. By contrast, the significantly negative estimate on risk-taking also shows with 20 PCs, but not beyond. This is in line with the observation from the regional analysis that risk-taking seems to be harder to predict from the Facebook data.

Italy

The first additional analysis for Italian regions shows that the significant positive association of patience with student achievement also holds for reading. Our main analysis in Section 3.4.1 focuses on math achievement, which is generally considered the most comparable subject across countries. Conversely, student reading outcomes are inherently language-specific, which makes them less suitable for cross-country analysis. We exploit the within-country nature and the richness of the INVALSI data to replicate our analysis using reading outcomes. Results in Table A3.5 show that a one SD increase in patience is associated with a 0.99-1.22 SD

increase in student reading achievement in the individual-level sample. At the regional level, a one SD increase in patience is associated with an increase of 0.71-0.91 SD in reading scores. The magnitude of the coefficients in reading is slightly smaller than in math but results clearly show in both subjects.

Results are also very robust across subsamples of waves and gender. The first two columns of Table A3.6 show that results do not depend on the year in which the assessment was conducted. This suggests that our estimates are not driven by the timing of the observation of the achievement data. Results also hold similarly for girls and boys, and the gender difference is not statistically significant (columns 3-4).[1]

In line with a leading role of cultural traits as a deep determinant of student achievement, results are stronger for native students than for migrant students. Results in Table A3.7 show that a one SD increase in patience is associated with a 1.42-1.58 SD increase in achievement for native students, a 0.75-0.91 SD increase in achievement for students with a second-generation migrant background, and a 0.56-0.89 SD increase in achievement for students with a first-generation migrant background. This pattern would be expected if it is indeed patience as a cultural trait that drives the achievement results, as the culture of the residence region is presumably less important for migrant students who have been less exposed to the regional culture.[2]

An additional robustness check ensures that results are not driven by student achievement in Trentino-Alto-Adige. In the INVALSI test of this region, only students in the autonomous municipalities of Bolzano and Trento are tested (see Section 3.3.1). This sampling in municipal areas only may bias our estimates, not least because Trentino-Alto-Adige is the Italian region with the highest estimated level of patience (see Section 3.2.3). Furthermore, we want to be sure that results are not driven by the Austrian history and the partially German-speaking population of the region. When omitting these municipalities from the analysis in Table A3.8, results are qualitatively the same and, if anything, slightly larger in magnitude.

We also perform an analysis of unobservable selection and coefficient stability proposed by Oster (2019). We compare our baseline models in Panel A of Table 3.2 to a restricted model without control variables. We follow the standard procedure and set $R_{max} = 1.3\tilde{R}$. The results in Table A3.9 imply that assuming an equal degree of selection between observables and unobservables, $\delta = 1$, the estimated bias-adjusted coefficient β^* for patience is between 1.487 and 1.705. In all cases, the bias-adjusted coefficient β^* is larger than our main estimates. The values δ for which $\beta = 0$ lie between -2.680 and -4.117. In all cases, these values are much larger than the standard cutoff $\delta = 1$. These results imply that the selection on unobservables would need to be more than 2.6 times larger than the selection on observables to push the coefficient of patience to 0.

[1] Reported results are based on Facebook-derived measures obtained with 4 PCs, but results are qualitatively the same with 7 and 10 PCs (not shown).

[2] Hanushek et al. (2022) find a similar pattern in their analysis of international student achievement.

Finally, we make use of the fact that Italy participated with a regionally representative sample in the international PISA test in 2012 to show that results hold equally well in this alternative achievement test. Intriguingly, the PISA results shown in Table A3.10 are very similar to the INVALSI results shown in Panel A of Table 3.2, indicating that a one SD increase in patience is associated with a 1.47-1.57 SD increase in the PISA math score.

United States

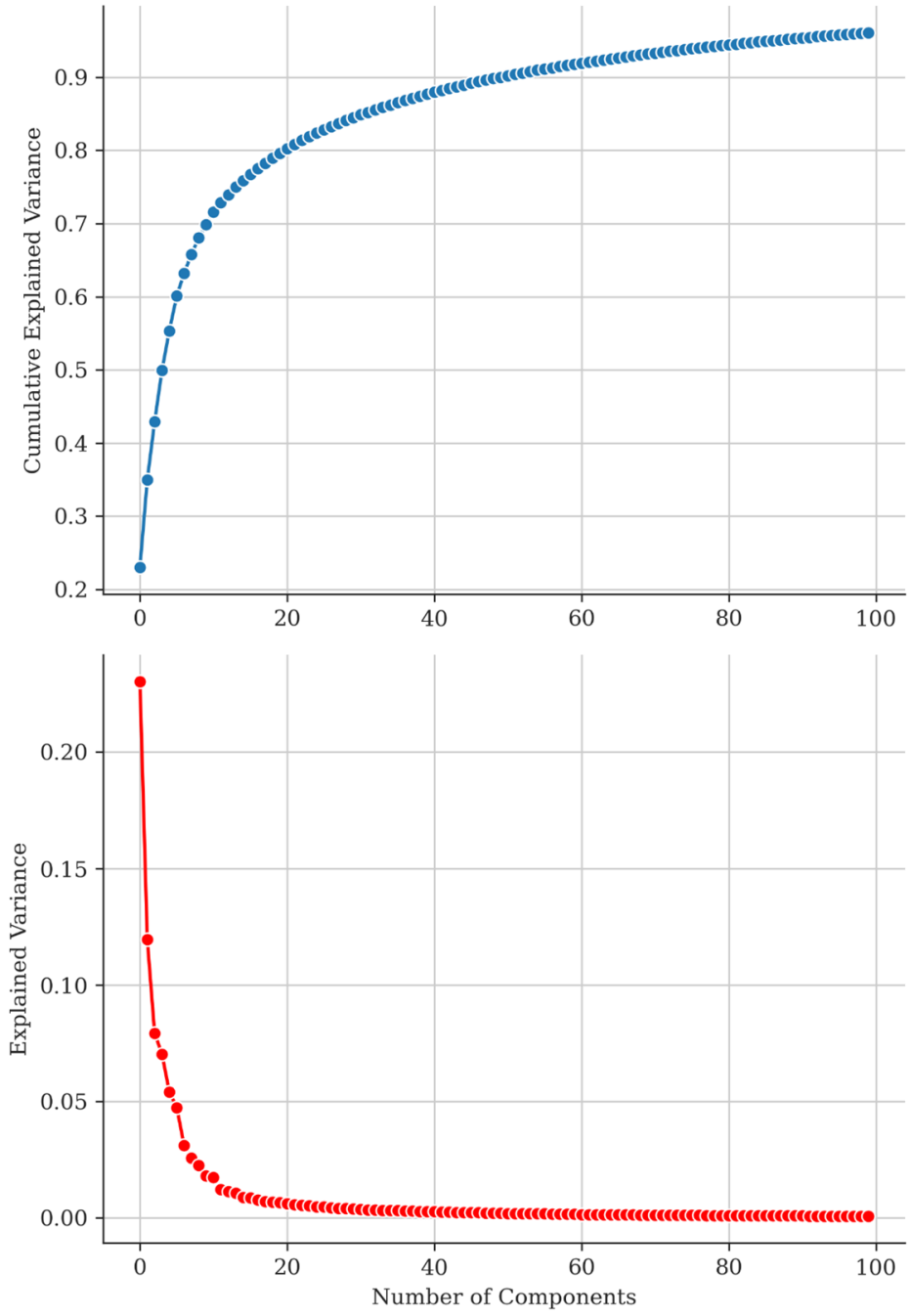
For the U.S. states, we first replicate the main results of the analysis in Section 3.4.2 using reading outcomes. The results reported in Table A3.11 closely mirror the findings for Italy: the magnitude of the coefficient of patience is slightly smaller compared the analysis of math achievement. A unit increase in patience is associated with an increase of 0.14-0.23 SD in reading achievement. Again, this analysis confirms that results do not depend on a particular subject.

We also check that results do not depend on the specific year in which student achievement is observed. Table A3.12 reports results using each wave of NAEP data – 2015, 2017, and 2019 – separately. Results are qualitatively the same for all analyzed waves. The magnitude of the patience coefficient tends to be smaller in the most recent wave, although not statistically significantly so. Overall, these results suggest that the findings do not depend on the specific year in which student test scores are observed.

Finally, the U.S. results are also similar across genders. Results in Table A3.13 show that patience is significantly positively associated with student achievement of both boys and girls. The coefficient estimates are somewhat larger for boys than for girls, but not significantly so, suggesting that results are qualitatively similar with respect to student gender.

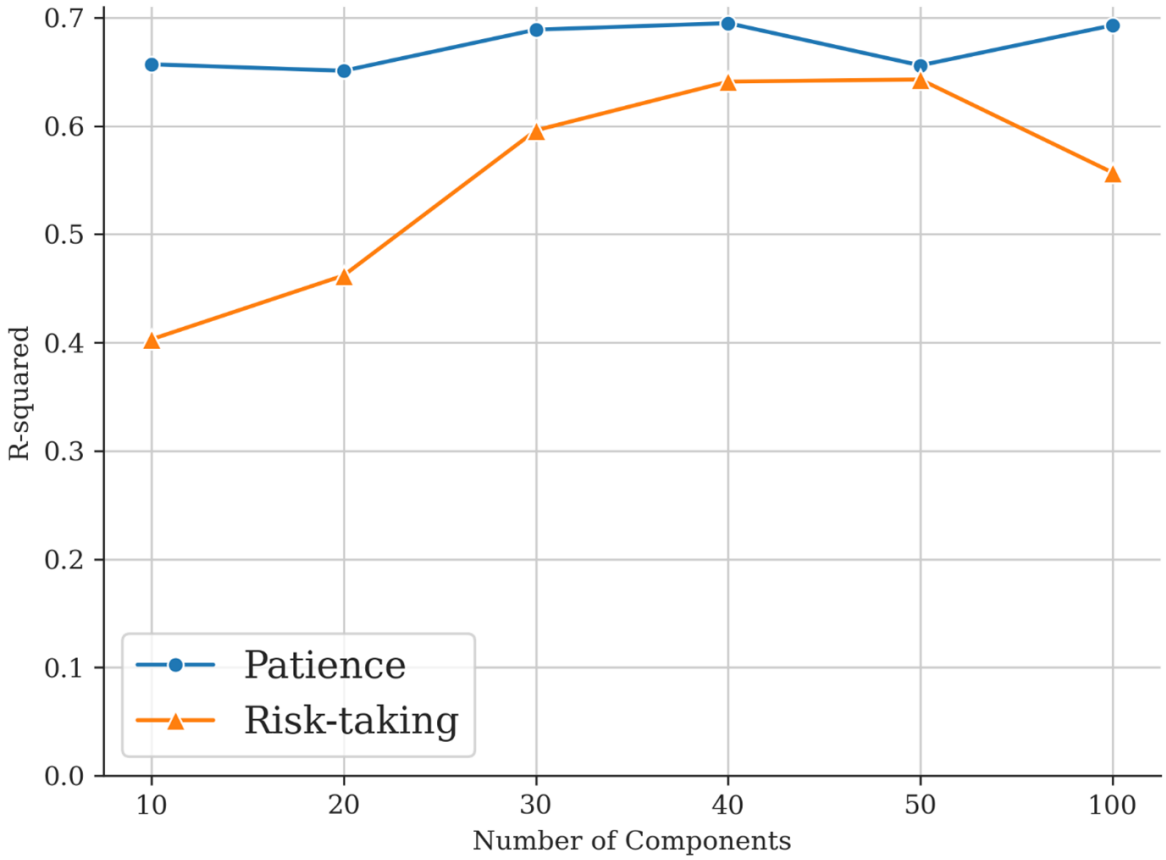
A3.2 Appendix Tables and Figures

Figure A3.1 : Variance in Facebook Interests Captured by PCs: International Sample



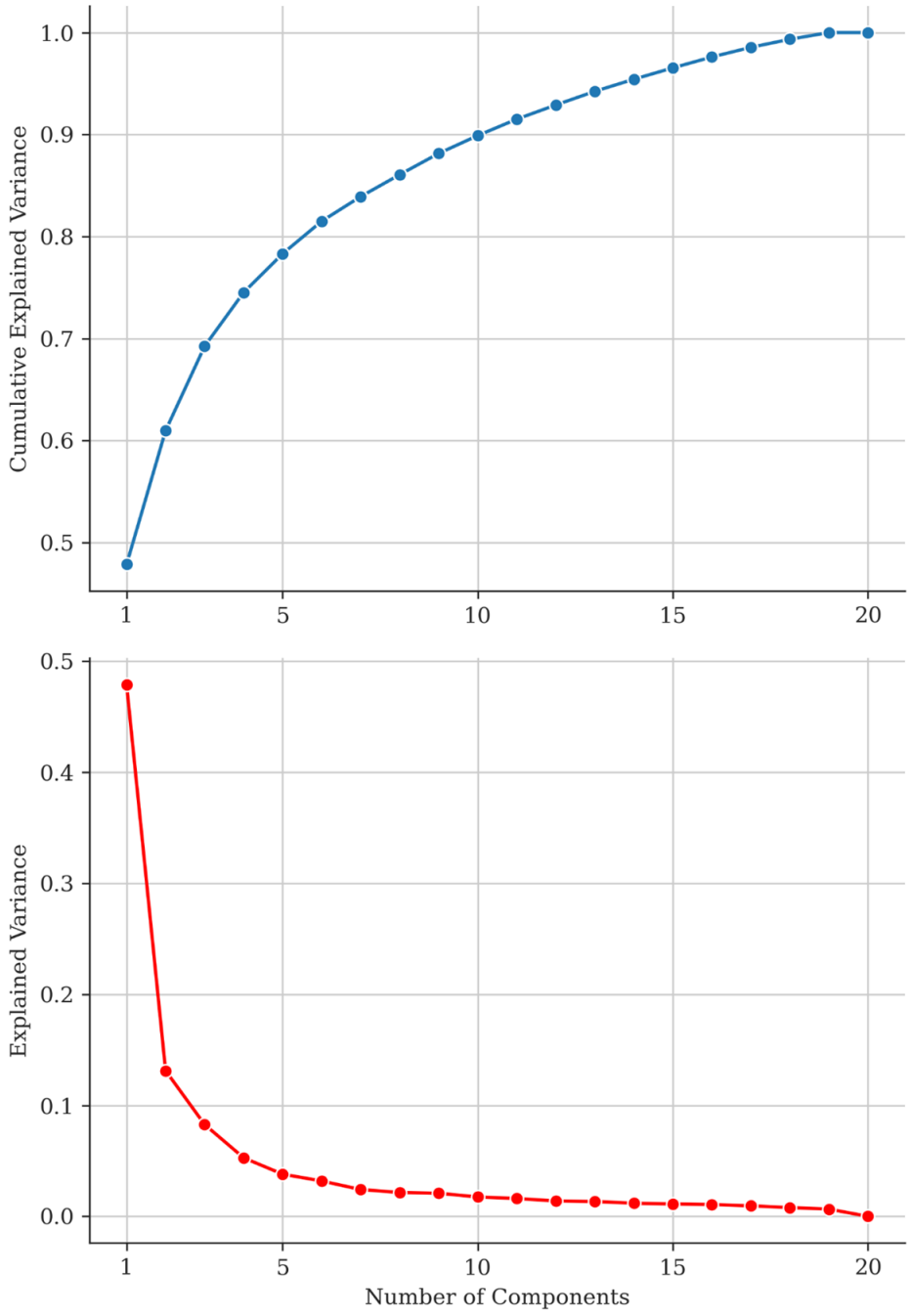
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the international sample, the bottom figure shows the variance captured by each component.

Figure A3.2 : Performance of GPS Prediction with Facebook Interests: International Sample



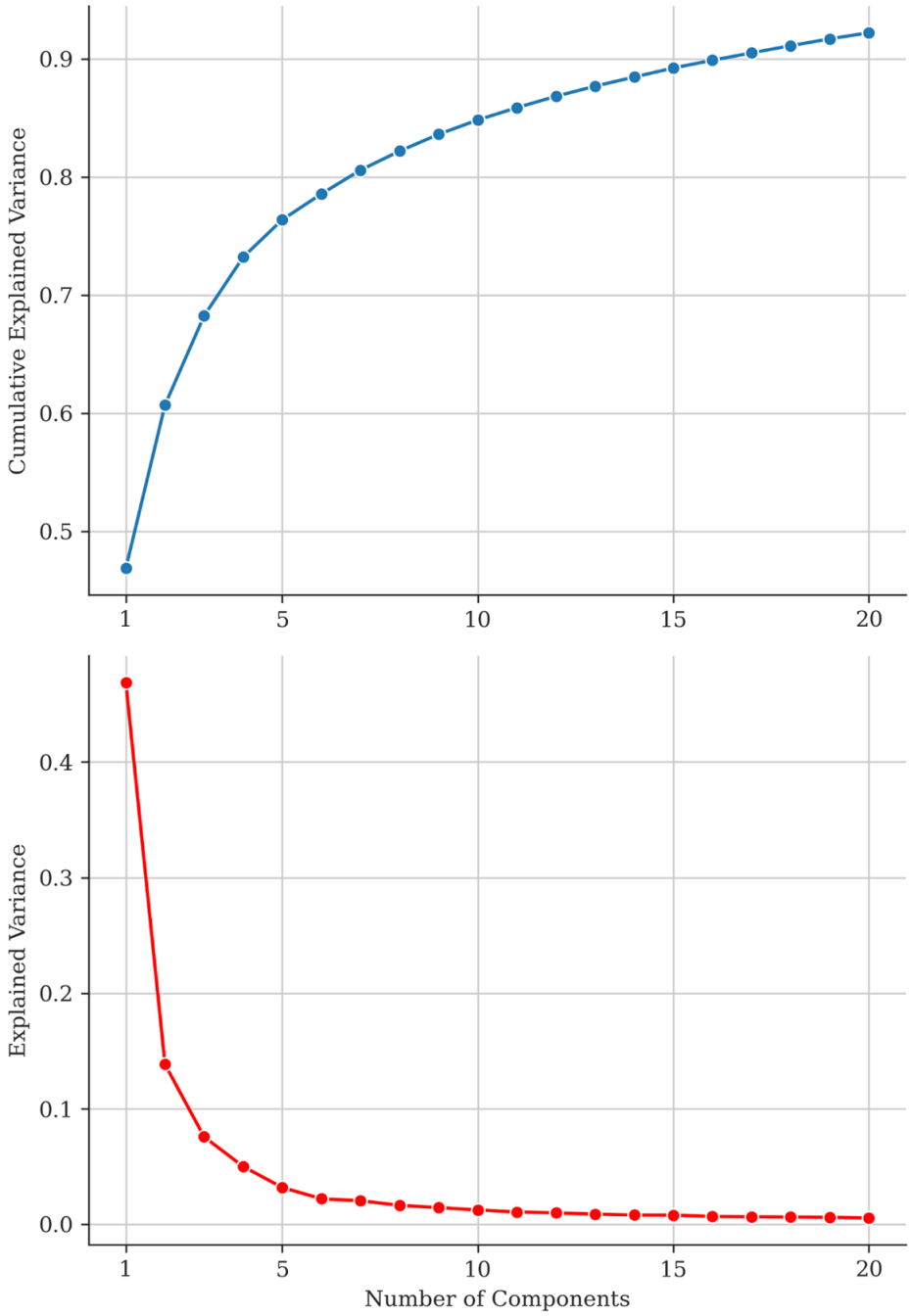
Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with PC loadings of country-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A3.3 : Variance in Facebook Interests Captured by PCs: Italian Regions



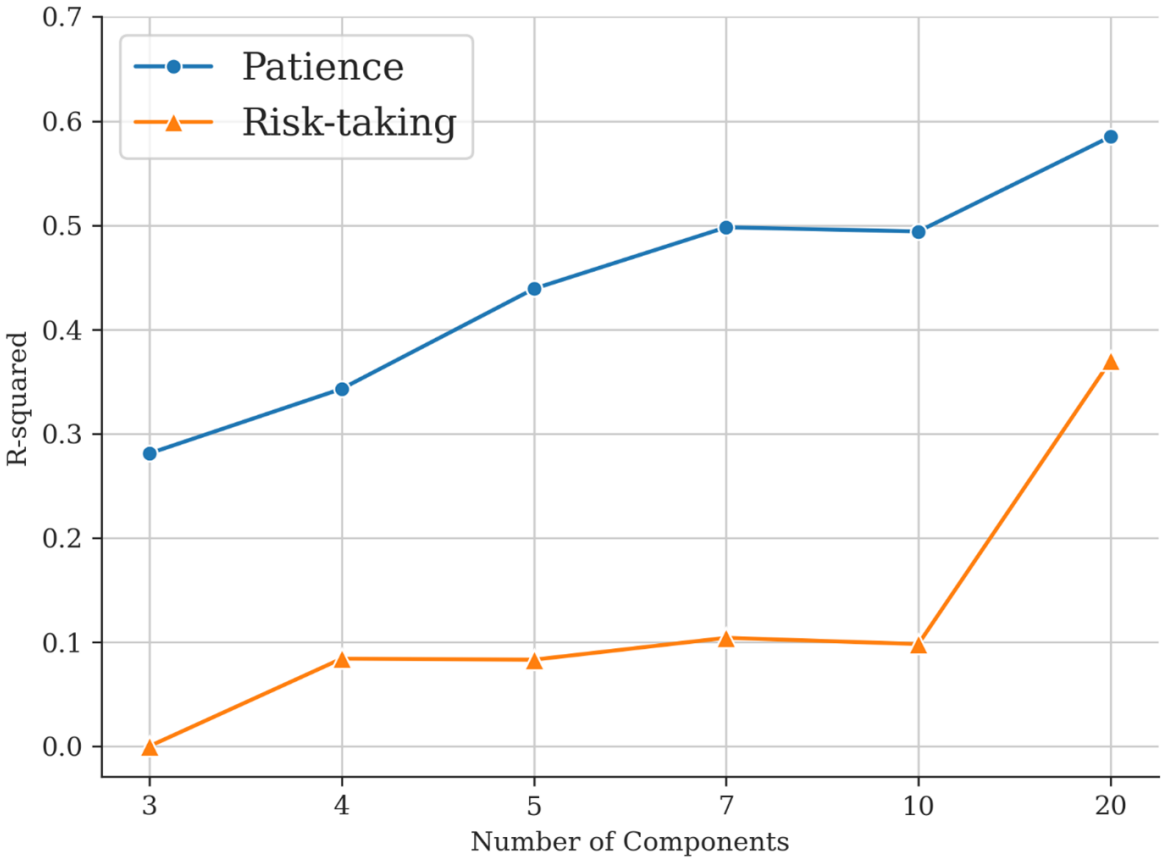
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the Italian regions, the bottom figure shows the variance captured by each component.

Figure A3.4 : Variance in Facebook Interests Captured by PCs: U.S. Regions



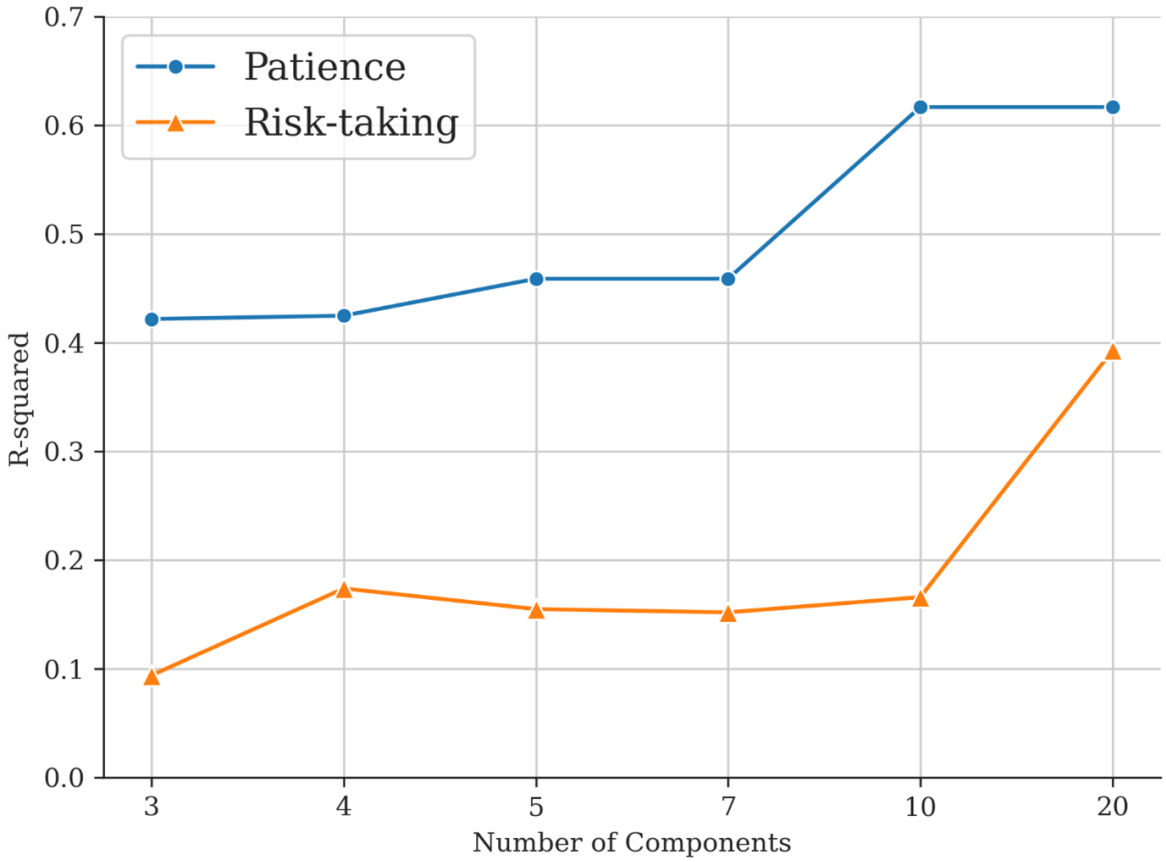
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the U.S. states, the bottom figure shows the variance captured by each component.

Figure A3.5 : Performance of GPS Prediction with Facebook Interests: PC Loadings from Italian Regions



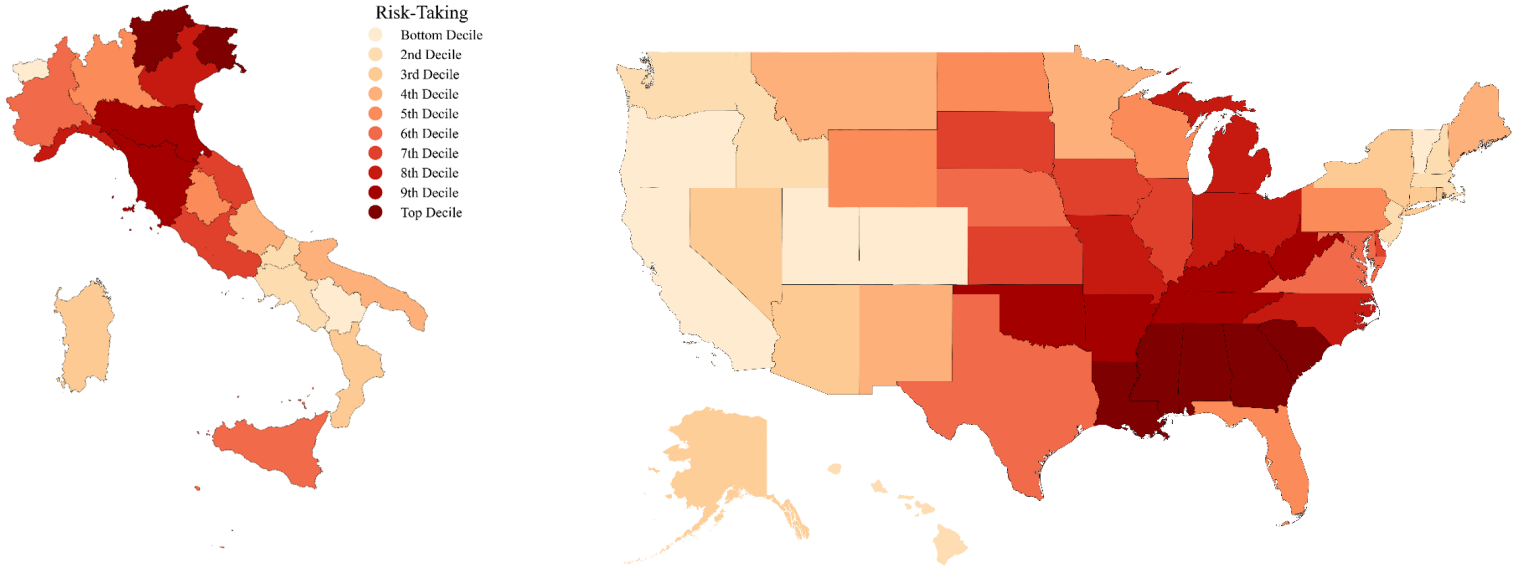
Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with the PC loadings of Italian-region-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A3.6 : Performance of GPS Prediction with Facebook Interests: PC Loadings from U.S. States



Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with PC loadings of U.S. state-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A3.7 : Patience, Risk-taking, and Student Achievement across Countries.



A. Italian Regions **B. U.S. States**

Notes: The figures show maps of the Facebook-derived measure of risk-taking obtained with 4 PCs for Italian regions (Panel A) and U.S. states (Panel B), respectively. Each color corresponds to a decile of the distribution of risk-taking within each country. Darker colors denote higher levels of risk-taking.

Table A3.1 : Countries in the Cross-Country Validation Exercise

	PISA countries			Training sample
	Only Facebook	Only GPS	Facebook and GPS	Facebook and GPS
	(1)	(2)	(3)	(4)
Afghanistan				x
Albania	x			
Algeria			x	x
Argentina			x	x
Australia			x	x
Austria			x	x
Azerbaijan	x			
Bangladesh				x
Belarus	x			
Belgium	x			
Bolivia				x
Bosnia and Herzegovina			x	x
Botswana				x
Brazil			x	x
Brunei Darussalam	x			
Bulgaria	x			
Cambodia				x
Cameroon				x
Canada			x	x
Chile			x	x
China				x
Colombia			x	x
Costa Rica			x	x
Croatia			x	x
Czech Republic			x	x
Denmark	x			
Dominican Republic	x			
Egypt				x
Estonia			x	x
Finland			x	x
France			x	x
Georgia			x	x
Germany			x	x
Ghana				x
Greece			x	x
Guatemala				x
Haiti				x
Hong Kong	x			
Hungary			x	x
Iceland	x			
India				x
Indonesia			x	x
Iraq				x
Ireland	x			
Israel			x	x
Italy			x	x
Japan			x	x
Jordan			x	x

(continued on next page)

Table A3.1 (continued)

	PISA countries			Training sample
	Only Facebook	Only GPS	Facebook and GPS	Facebook and GPS
	(1)	(2)	(3)	(4)
Kazakhstan			x	x
Kenya				x
Korea			x	x
Kyrgyzstan	x			
Latvia	x			
Lebanon	x			
Liechtenstein	x			
Lithuania			x	x
Luxembourg	x			
Macao	x			
Malawi				x
Malaysia	x			
Malta	x			
Mauritius	x			
Mexico			x	x
Moldova			x	x
Montenegro	x			
Morocco			x	x
Netherlands			x	x
New Zealand	x			
Nicaragua				x
Nigeria				x
North Macedonia	x			
Norway	x			
Pakistan				x
Panama	x			
Peru			x	x
Philippines			x	x
Poland			x	x
Portugal			x	x
Qatar	x			
Romania			x	x
Russia		x		
Rwanda				x
Saudi Arabia			x	x
Serbia			x	x
Singapore	x			
Slovakia	x			
Slovenia	x			
South Africa				x
Spain			x	x
Sri Lanka				x
Suriname				x
Sweden			x	x
Switzerland			x	x
Tanzania				x
Thailand			x	x
Trinidad and Tobago	x			

(continued on next page)

Table A3.1 (continued)

	PISA countries			Training sample
	Only Facebook	Only GPS	Facebook and GPS	Facebook and GPS
	(1)	(2)	(3)	(4)
Tunisia	x			
Turkey			x	x
Uganda				x
Ukraine			x	x
United Arab Emirates			x	x
United Kingdom			x	x
United States			x	x
Uruguay	x			
Venezuela				x
Vietnam			x	x
Zimbabwe				x
Total: 107 countries	32	1	48	74

Notes: Sample of countries: Col. 1-3: countries included in the cross-country validation exercise (Panel A of Table 3.1). Col. 4: countries included in training the machine learning model. Country names are as reported in PISA codebooks or Facebook/GPS data and do not represent any political views of the authors.

Table A3.2 : Countries in the Migrant Analysis

	GPS/Facebook country of origin			PISA destination country	
	Only GPS	Only Facebook	Both	GPS analysis	Facebook analysis
	(1)	(2)	(3)	(4)	(5)
Afghanistan			x		
Albania		x			
Algeria					
Argentina			x	x	x
Armenia		x			
Australia			x	x	x
Austria			x	x	x
Azerbaijan		x			
Bangladesh			x		
Belarus		x		x	x
Belgium		x		x	x
Bolivia			x		
Bosnia and Herzegovina			x	x	x
Brazil			x		
Brunei Darussalam				x	x
Bulgaria		x			
Cape Verde		x			
Canada			x	x	x
Chile			x		
China			x		
Colombia			x		
Costa Rica				x	x
Croatia			x	x	x
Czech Republic			x	x	x
Denmark		x		x	x
Dominican Republic		x		x	x
Egypt			x		
Estonia			x		
Ethiopia		x			
Fiji		x			
Finland			x	x	x
France			x		
Georgia			x		x
Germany			x	x	x
Greece			x		x
Haiti			x		
Hong Kong				x	x
Hungary			x		
Iceland		x			
India			x		
Indonesia			x	x	x
Iran	x				
Iraq			x		
Ireland		x		x	x
Israel				x	x
Italy			x		
Japan					
Jordan			x	x	x

(continued on next page)

Table A3.2 (continued)

	GPS/Facebook country of origin			PISA destination country	
	Only GPS	Only Facebook	Both	GPS analysis	Facebook analysis
	(1)	(2)	(3)	(4)	(5)
Kazakhstan			x		
Kuwait		x			
Kyrgyzstan					x
Latvia				x	x
Lebanon		x			
Libya		x			
Liechtenstein		x		x	x
Lithuania			x		
Luxembourg				x	x
Macao		x		x	x
Malaysia		x			
Mauritius				x	x
Mexico				x	x
Moldova			x	x	x
Montenegro		x		x	x
Morocco			x	x	x
Netherlands			x	x	x
New Zealand		x		x	x
Nicaragua			x		
Nigeria			x		
North Macedonia				x	x
Norway		x		x	x
Pakistan			x		
Palestine		x			
Panama		x		x	x
Paraguay		x			
Peru					
Philippines			x	x	x
Poland			x		
Portugal			x	x	x
Qatar		x		x	x
Romania			x		
Russia	x				
Samoa		x			
Saudi Arabia			x	x	x
Serbia			x		x
Singapore		x			
Slovakia		x		x	x
Slovenia		x		x	x
Somalia		x			
South Africa			x		
South Korea			x	x	x
Spain			x		
Suriname			x		
Sweden			x		
Switzerland			x	x	x
Tajikistan		x			
Thailand			x		

(continued on next page)

Table A3.2 (continued)

	GPS/Facebook country of origin			PISA destination country	
	Only GPS	Only Facebook	Both	GPS analysis	Facebook analysis
	(1)	(2)	(3)	(4)	(5)
Tonga		x			
Turkey			x	x	x
Ukraine			x	x	x
United Arab Emirates			x		
United Kingdom			x	x	x
United States			x		
Uruguay		x		x	x
Uzbekistan		x			
Venezuela			x		
Vietnam			x		
Yemen		x			
Zambia		x			
Total: 108 countries	2	37	56	46	50

Notes: Sample of countries that serve as countries of origin (col. 1-3) or destination countries (col. 4-5) in the migrant analysis (Panel B of Table 3.1). Country names are as reported in PISA codebooks or Facebook/GPS data and do not represent any political views of the authors.

Table A3.3 : Validation of Cross-Country Analysis: Different Numbers of Principal Components (PCs)

	20 PCs	30 PCs	40 PCs	50 PCs
	(1)	(2)	(3)	(4)
A. Original country sample (GPS countries)				
Patience	1.598***	1.588***	1.601***	1.610***
	(0.132)	(0.140)	(0.139)	(0.140)
Risk-taking	-1.598***	-0.883***	-0.898***	-1.004***
	(0.452)	(0.316)	(0.308)	(0.276)
Control variables	Yes	Yes	Yes	Yes
Observations	1,954,840	1,954,840	1,954,840	1,954,840
Residence countries	48	48	48	48
R^2	0.207	0.195	0.197	0.202
B. Extended country sample (all Facebook countries)				
Patience	1.641***	1.598***	1.607***	1.597***
	(0.121)	(0.126)	(0.129)	(0.130)
Risk-taking	-1.640***	-1.265***	-1.160***	-1.126***
	(0.336)	(0.285)	(0.263)	(0.229)
Control variables	Yes	Yes	Yes	Yes
Observations	2,660,408	2,660,408	2,660,408	2,660,408
Residence countries	80	80	80	80
R^2	0.205	0.203	0.200	0.199

Notes: Dependent variable: PISA math test score in all PISA waves 2000-2018. Least squares regressions weighted by students' sampling probability. Control variables: student gender, age, and migration status; imputation dummies; and wave fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; own elaboration of Facebook data.

Table A3.4 : Validation of Migrant Analysis: Different Numbers of Principal Components (PCs)

	20 PCs	30 PCs	40 PCs	50 PCs
	(1)	(2)	(3)	(4)
A. Original sample (GPS countries of origin)				
Patience	0.783***	0.876***	0.885***	0.875***
	(0.193)	(0.197)	(0.192)	(0.216)
Risk-taking	-0.676**	0.008	0.087	0.156
	(0.306)	(0.367)	(0.322)	(0.371)
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	78,403	78,403	78,403	78,403
Countries of origin	56	56	56	56
Residence countries	46	46	46	46
R^2	0.271	0.271	0.272	0.270
B. Extended sample (all Facebook countries of origin)				
Patience	0.838***	1.027***	1.033***	0.995***
	(0.211)	(0.198)	(0.191)	(0.211)
Risk-taking	-1.155***	-0.067	0.064	0.154
	(0.422)	(0.357)	(0.297)	(0.341)
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	90,983	90,983	90,983	90,983
Countries of origin	93	93	93	93
Residence countries	50	50	50	50
R^2	0.295	0.294	0.294	0.291

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions, including 180 fixed effects for each residence-country by wave cell. Sample: students with both parents not born in the country where the student attends school. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; own elaboration of Facebook data.

Table A3.5 : Patience and Reading Achievement: Analysis of Italian Regions

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
A. Individual level			
Patience	1.218***	0.986***	1.050***
	(0.201)	(0.123)	(0.128)
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	59,441	59,441	59,441
Regions	20	20	20
R^2	0.105	0.110	0.110
B. Regional level			
Patience	0.905***	0.716***	0.762***
	(0.177)	(0.094)	(0.098)
Wave fixed effects	Yes	Yes	Yes
Observations	42	42	42
Regions	20	20	20
R^2	0.496	0.617	0.625

Notes: Dependent variable: INVALSI 8th-grade reading test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI reading achievement test, 2017-2019; own elaboration of Facebook data.

Table A3.6 : Patience and Math Achievement: Analysis of Italian Regions by Subgroups

	2018	2019	Males	Females
	(1)	(2)	(3)	(4)
A. Individual level				
Patience (4 PCs)	1.588***	1.422***	1.579***	1.427***
	(0.191)	(0.217)	(0.211)	(0.198)
Control variables	Yes	Yes	Yes	Yes
Wave fixed effects	No	No	Yes	Yes
Observations	29,359	29,675	30,530	28,504
Regions	20	20	20	20
R^2	0.095	0.089	0.097	0.082
B. Regional level				
Patience (4 PCs)	1.331***	1.161***	1.305***	1.185***
	(0.221)	(0.241)	(0.226)	(0.227)
Wave fixed effects	No	No	Yes	Yes
Observations	21	21	42	42
Regions	20	20	20	20
R^2	0.693	0.668	0.682	0.657

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Patience measure computed with 4 principal components (PCs). Regressions include the risk-taking measure computed with 4 PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI reading achievement test, 2017-2019; own elaboration of Facebook data.

Table A3.7 : Patience and Math Achievement: Analysis of Italian Regions by Migrant Status

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
A. Native students			
Patience	1.581***	1.423***	1.514***
	(0.188)	(0.115)	(0.118)
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	51,691	51,691	51,691
Regions	20	20	20
R^2	0.084	0.091	0.091
B. Second-generation migrant students			
Patience	0.909***	0.748***	0.820***
	(0.237)	(0.215)	(0.220)
Wave fixed effects	Yes	Yes	Yes
Observations	3,572	3,572	3,572
Regions	20	20	20
R^2	0.033	0.035	0.035
C. First-generation migrant students			
Patience	0.565**	0.842***	0.893***
	(0.235)	(0.112)	(0.124)
Wave fixed effects	Yes	Yes	Yes
Observations	1,719	1,719	1,719
Regions	20	20	20
R^2	0.079	0.083	0.083

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: student. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables: student gender and age; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A3.8 : Patience and Math Achievement: Analysis of Italian Regions Excluding Trentino-Alto-Adige

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
A. Individual level			
Patience	1.717***	1.412***	1.520***
	(0.158)	(0.122)	(0.124)
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	55,437	55,437	55,437
Regions	19	19	19
R^2	0.095	0.098	0.098
B. Regional level			
Patience	1.462***	1.220***	1.314***
	(0.171)	(0.094)	(0.097)
Wave fixed effects	Yes	Yes	Yes
Observations	38	38	38
Regions	19	19	19
R^2	0.783	0.835	0.846

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Students in the autonomous municipalities of Trento and Bolzano are dropped from the estimation sample. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A3.9 : Analysis of Unobservable Selection and Coefficient Stability following Oster (2019): Analysis of Italian Regions

	4 PCs		7 PCs		10 PCs	
	Restricted	Extended	Restricted	Extended	Restricted	Extended
	(1)	(2)	(3)	(4)	(5)	(6)
Patience	1.252***	1.505***	1.136***	1.350***	1.208***	1.437***
	(0.210)	(0.197)	(0.122)	(0.114)	(0.129)	(0.117)
Control variables	No	Yes	No	Yes	No	Yes
Wave fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	59,034	59,034	59,034	59,034	59,034	59,034
Regions	20	20	20	20	20	20
R^2	0.043	0.092	0.049	0.099	0.050	0.099
Oster (2019) diagnostics						
Bound β^* for $\delta = 1$	1.705		1.487		1.581	
δ to match $\beta = 0$	-4.117		-2.687		-2.680	

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: student. Students in the autonomous municipalities of Trento and Bolzano are dropped from the estimation sample. Patience measure computed with number of principal components (PCs) indicated in column header. Regressions include the risk-taking measure computed with the equivalent number of PCs. Odd columns: restricted model with wave fixed effects. Even columns: baseline models with wave fixed effects, student gender, age, and migration status; imputation dummies. Oster statistics computed using $R_{max} = 1.3\tilde{R}$, where $\tilde{R}$ denotes the R^2 reported in even columns. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A3.10 : Patience and Math Achievement: Analysis of Italian Regions using PISA 2012 Data

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
Patience	1.484***	1.473***	1.570***
	(0.264)	(0.132)	(0.138)
Control variables	Yes	Yes	Yes
Observations	31,073	31,073	31,073
States	20	20	20
R^2	0.106	0.113	0.113

Notes: Dependent variable: PISA 2012 math test score. Least squares regressions. Unit of observation: student. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Control variables: student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA student achievement test, 2012; own elaboration of Facebook data.

Table A3.11 : Patience and Reading Achievement: Analysis of U.S. States

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
Patience	0.228***	0.141*	0.227**
	(0.074)	(0.077)	(0.103)
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.385	0.375	0.396

Notes: Dependent variable: NAEP 8th-grade reading test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table A3.12 : Patience and Math Achievement: Analysis of U.S. States by Wave

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
A. 2015			
Patience	0.335***	0.194**	0.346***
	(0.081)	(0.082)	(0.119)
States	51	51	51
R^2	0.426	0.410	0.430
B. 2017			
Patience	0.309***	0.179**	0.290**
	(0.084)	(0.085)	(0.125)
States	51	51	51
R^2	0.373	0.360	0.372
C. 2019			
Patience	0.235***	0.142*	0.228*
	(0.077)	(0.077)	(0.114)
States	51	51	51
R^2	0.277	0.267	0.278

Notes: Dependent variable: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table A3.13 : Patience and Math Achievement: Analysis of U.S. States by Gender

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Males			
Patience	0.322***	0.194*	0.305**
	(0.101)	(0.108)	(0.147)
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.388	0.377	0.385
B. Females			
Patience	0.263***	0.147*	0.258**
	(0.079)	(0.086)	(0.119)
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.319	0.304	0.321

Notes: Dependent variable: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

4 Good or Bad News First? The Effect of Feedback Order on Motivation and Performance*

4.1 Introduction

Feedback is crucial in any learning environment, from education to managerial contexts as well as personal interactions. Learning about the assessment of one's performance is an important channel for analyzing and potentially changing behavior (Ammons, 1956; Kluger and DeNisi, 1996; Villeval, 2022). However, there is little consensus on how to 'best' provide feedback. Various theories and practices have emerged, from the famous 'feedback sandwich' that embeds one corrective feedback into two positive feedback elements in a managerial context (see e.g. Davies and Jacobs (1985)) to the mostly criticizing feedback in the academic context (Allgood et al., 2019; Dupas et al., 2021).

A potentially meaningful element of feedback that is easy to adapt is the ordering of feedback parts. Classic economic theories emphasize the information content of feedback which helps to reduce uncertainty, e.g. about one's own performance. Individuals adjust their performance beliefs which is mostly assumed to follow Bayes' rule: prior and new information are combined to attain an assessment of one's performance. If that was the only aspect of feedback, the ordering should not matter as long as the information content of the differently ordered feedback elements is the same. Instead, theories from psychology support potential effects of feedback ordering due to emotional reactions to feedback (Choi et al., 2018) as well as attributing the feedback to the self rather than the performance on the task (Kluger and DeNisi, 1996).

In this paper, I conduct a field experiment to study whether the ordering of positive and negative feedback elements matters for the motivation and performance of university students. Each treated student receives one positive and one negative feedback part on their performance in exam practice questions. The feedback refers to subtopics of the practice questions such that the positive (negative) feedback element informs students about their best (worst) subtopic, independently of their ranking compared to other students. Besides the respective topic, the wording of both feedback elements is identical for all students. I randomly vary the ordering of the two feedback elements across the treatment groups to test whether the order affects students' motivation to study for the respective exam and their performance in subsequent practice questions and the exam.

* This chapter is based on the job market paper 'Good or Bad News First? The Effect of Feedback Order on Motivation and Performance', mimeo. The project was funded through an Add-On Fellowship from the Joachim Herz Foundation. IRB approval was obtained from the Ethics Committee of LMU Munich (Project 2021-19) and the experiment was pre-registered in the AEA RCT Registry (AEARCTR-0008953).

I focus on motivation and performance due to their importance for later labor market outcomes. A large literature shows that higher educational performance is associated with higher wages and lower unemployment (Zax and Rees, 2002; Hanushek et al., 2015). Similarly, evidence from education science shows how intrinsic motivation for a specific activity is negatively related to drop out from (higher) education (Gillet et al., 2012; Cabus and De Witte, 2016; Rump et al., 2017). In economics, seminal work from Bénabou and Tirole (2002) contextualizes how self-confidence and motivation interact. Through motivation, self-confidence can improve individuals' perseverance and thus compensate for imperfect willpower.

I find that first receiving positive feedback leads to higher post-feedback motivation to study for the exam compared to receiving negative feedback first. To understand how this effect emerges over the course of the experiment, I compare both treatment groups to the pre-treatment average motivation of a control group that does not receive any feedback on the practice questions. The positive effect of the *positive-negative* (POSNEG) ordering compared to *negative-positive* (NEGPOS) is driven by a drop in motivation for the NEGPOS group after the negative feedback part that persists after they have also received the positive feedback element. This drop in motivation is not observed for participants in the POSNEG group who receive the negative feedback as a second element, which points to a shielding effect of the positive feedback when receiving it first. For either treatment group, positive feedback does not affect motivation compared to the control group.

The effect of feedback ordering on motivation in the two treatment groups cannot be explained by incorrect belief updating: there is no significant effect on post-treatment beliefs about performance in the practice questions. Compared to the control group, performance beliefs follow closely the patterns from motivated beliefs where positive feedback is over-weighted compared to negative feedback, although not differentially by feedback order. Instead, overall feelings about the feedback elicited one day after the exam are much better for students from the POSNEG group compared to the NEGPOS group. This suggests that the effects on motivation might be driven by emotional responses to the feedback received.

Furthermore, I find that students strongly react to the feedback topics and adjust their study content accordingly. Over the course of the experiment, students are asked multiple times which chapters of the respective course they plan to focus on in the remaining study time before the exam. From that, I construct an indicator of whether the respective topic they receive positive or negative feedback on was part of that individual study list before treatment, between the two feedback elements, and after both feedback elements. For both treatment groups, a clear pattern emerges compared to the pre-treatment allocation of the control group: students persistently remove (add) course topics from (to) their study list as soon as they receive the respective positive (negative) feedback on them. This happens more strongly for the negative feedback topic than for the positive feedback topic and students are slightly more cautious about removing the positive feedback topic from their study list if they have received negative feedback first. These findings imply that students understand the content of the feedback and consider it relevant. The pattern can still be observed when

asking students about their actual study content one day after the exam, although slightly weaker in magnitude. No effect can be found on study hours.

Finally, there is no significant overall effect of feedback ordering on performance in a second set of practice questions and in the final exam. This might be related to a lack of statistical power in detecting reasonably-sized effects for such an outcome, but might also reflect other hurdles in translating the effects of feedback ordering on motivation to students' exam performance. When having a closer look at the exam content, I find that there seems to be a positive effect of the *positive-negative* ordering on exam grades for those students who encounter their negative feedback topic in the exam. For these students, it might be especially important to have avoided a drop in study motivation from negative feedback when handling the presence of the negative feedback topic in the exam.

The results of feedback ordering on motivation suggest that the information content of feedback might not be the only mechanism in place. In this setting, the informativeness of the feedback is the same for the two treatment groups since students in both receive feedback on a topic they performed best in as well as worst. The resulting treatment effects on motivation and post-exam feelings about feedback imply a more emotional or impulsive reaction that leads to a different perception of positive and negative feedback elements depending on their ordering. This is supported by the psychological literature on how feedback can cause emotional reactions (see e.g. Erickson et al. (2021)) that then in turn can shape (cognitive) behavior (see e.g. Zadra and Clore (2011), Baumeister et al. (2007), and Tyng et al. (2017)). The findings hence add a new angle to the interpretation of feedback in economics, suggesting that individuals react to more than just the information content of feedback in a higher education setting.

The way feedback is provided in this intervention has a series of advantages which might nicely complement other types of feedback. Highlighting individual strengths and weaknesses can be crucial for personal development, especially at young ages. It might help students develop an assessment of their capacities that is independent of others, which in turn can make it easier to adapt to new settings with different peers. More specifically, informing students about the course topics they are currently performing best and worst in when studying for an exam, could be helpful in terms of study efficiency. If we consider studying for an exam to be an optimization problem of how to best allocate study time under a constrained time budget, advice on where to invest more time might be as useful as information on where to take that time away. Similarly, the within-person feedback used in this study can reduce inequality in the direction of feedback received by students. Relative performance feedback usually implies that students in the lower part of the distribution hardly ever receive positive feedback. Most likely though, these students also have strengths: they might either be low-performers just with respect to a specific peer group or have relative strengths in certain areas versus others. Hence, information about a wider range of their strengths and weaknesses instead of just their relative position compared to a specific peer group can help especially lower achievers

to make better decisions for future life choices. Furthermore, this within-person feedback can be an alternative solution when no clear reference group can be identified.

This paper mainly contributes to three strands of the literature. First, it adds to the literature on feedback interventions in education. A series of papers has investigated the effect of absolute and relative performance feedback in a university setting. Most of these find positive effects of providing feedback, especially on those receiving positive feedback (see e.g. Bandiera et al. (2015) in the UK, Brade et al. (2022) in Germany, Kajitani et al. (2020) in Japan, and Dobrescu et al. (2021) in Australia). Instead, Azmat et al. (2019) find negative short-run effects of feedback on college students' performance in Spain that are driven by those who initially underestimate their performance, i.e. receive positive relative performance feedback. Similarly, papers focusing on secondary education have largely found positive effects of providing relative performance feedback on subsequent performance (Azmat and Iriberri, 2010; Fischer and Wagner, 2018; Goulas and Megalokonomou, 2021). An exception to this is the paper by Bursztyn and Jensen (2015) where publicly revealing top performances makes the best students reduce their effort to avoid being exposed to their peer group. Lastly, a smaller number of studies has looked at feedback in primary education. Muis et al. (2015) find that performance feedback in kindergarten increases performance but reduces enjoyment, whereas Hermes et al. (2021) see positive effects of informing low achievers about performance improvements on motivation, effort, and performance without any negative effects on high achievers. The contribution of this paper to this literature is based on the type of feedback I provide in the intervention: by moving away from peer-related relative performance feedback which is highly dependent on the respective reference group, I can study the effects of the ordering of positive *and* negative feedback elements as well as the effects of positive (negative) feedback on low (high) achievers while still providing truthful and informative feedback. Furthermore, I provide feedback on specific topics within an overall performance which is a novel approach with respect to the existing literature.

The second related literature about feedback is the one from behavioral economics. Eil and Rao (2011) show how feedback on ego-relevant dimensions such as intelligence and beauty is incorporated differently by individuals depending on whether it is positive or negative compared to their prior assessment. Möbius et al., 2022 find similar evidence of such motivated beliefs for undergraduate students, again especially in ego-relevant dimensions. Both studies show that individuals tend to over-weight positive feedback relative to negative feedback. In an emerging part of this literature, Zimmermann (2020) has highlighted the dynamic aspect of motivated beliefs. His findings confirm that individuals react more to positive feedback which becomes even more evident after some time passes. Following up on this approach, Coffman et al. (2021) document large and persistent gender gaps in beliefs and choices and show that women react more strongly to negative feedback compared to men. I add to this literature in two ways: first, my experiment combines the study of the effects of the respective feedback elements themselves and their ordering with a dynamic perspective on how effects of feedback can potentially be explained by the evolution of individuals' reactions to the

separate elements. Second, moving the setting from a laboratory (as has been used in most of the previous studies in this literature) to a field setting strengthens the conclusions that can potentially be drawn from the analysis. This is especially true since education is one of the crucial periods not only for human capital accumulation leading to far-reaching later life decisions but also for the development of one's personality.

Lastly, the paper builds on previous work from psychology. Feedback has been a topic of interest in this literature for a very long time, reaching back to first important theoretical contributions from Thorndike (1913, 1927). In the following decades, potential mechanisms of individuals' reactions to feedback have been further conceptualized in e.g. Ammons (1956) and Ilgen et al. (1979), while first small empirical studies have been conducted to understand how e.g. the ordering of feedback elements is interpreted by individuals (Jacobs et al., 1973; Schaible and Jacobs, 1975; Davies and Jacobs, 1985). Most of these studies were conducted on very small and/or selected samples or only included a subset of potential feedback orderings such that the resulting evidence is only partially conclusive. A further milestone in this literature was the introduction of the so-called feedback intervention theory by Kluger and DeNisi (1996) that highlights how the effectiveness of feedback decreases when it is given or perceived as closer to the self than the task. More recently, empirical studies have focused on the feedback sandwich, with mixed results. For example, Slowiak and Lakowske (2017) and Henley and DiGennaro Reed (2015) show that there was no differential impact of different sandwich orderings on performance, but participants rather preferred the non-sandwich version with corrective-positive-positive feedback. Furthermore, Choi et al. (2018) provide evidence on feedback ordering in a work-related lab experiment and show how emotional responses can play a role in reactions to feedback. Whereas most of these approaches looking at the feedback sandwich cannot isolate the effect of the presence of more than one positive or negative feedback element, my study has a clean design only using one positive and one negative element, in a randomly varying order. Furthermore, it brings the analysis of feedback ordering to a field setting in education where feedback is especially relevant since students are constantly accompanied by different types of feedback givers in their adolescent development.

The remainder of the paper is structured as follows: section 4.2 details the experimental design whereas section 4.3 gives an overview of the sample and descriptive statistics. Section 4.4 shows the main experimental results, section 4.5 concludes.

4.2 Experimental Design

The field experiment was implemented in the setting of university courses that undergraduate Business and Economics students have to take at the University of Munich. I set up an additional exam preparation that students could sign up for. Students from two courses par-

ticipated in February 2022 which was the exam period of the Winter Semester 2021/2022.[1] All information and questionnaires were presented in German according to the course language.[2]

The flow chart in figure 4.1 shows the overall sequence of the study design. The following subsections explain each of the parts in detail.

4.2.1 Sign-up

The opportunity to participate in the study was announced both during (online) lectures as well as via email three to four weeks before the exam. Students were informed that they had the chance to practice their knowledge in exam-type questions and that they would receive personalized feedback on their performance. After the deadline for signing up (10 days before the exam), students received an email with the outline of the exam preparation. This email included the respective dates for the additional exam preparation questions (*seven* and *three* days before the exam), for the feedback I would provide (*three* days before the exam), and for a short post-exam questionnaire (*one* day after the exam). It was specified that only those who participated in all three questionnaires had the chance of winning a 50 Euro voucher to be chosen from the platforms Amazon, Avocadostore, and Netflix. Around 20% of participants received such a voucher in the end. Finally, students were informed that they would be contacted again once they had received the grades for the respective exam and that each student who would then report the grade together with some proof (e.g. a screenshot) would receive an additional 10 Euro voucher to be chosen from the same platforms.[3]

4.2.2 First Set of Practice Questions

Seven days before the exam, students received an email with a link to the first questionnaire including the first exam preparation questions. Prior to the practice questions, students answered a series of questions on some background characteristics as well as personality traits. The former contained questions on age, gender, semester, field of study, their high school GPA and last high-school math grade as a proxy for ability, their parents' highest education and employment status as a proxy for socio-economic status as well as a question on whether students and their parents were born in Germany as a proxy for culture. The

[1] Both courses were related to basic methods and concepts that are taught in the first semesters of a typical undergraduate Business and Economics program. Exams were concentrated in a period of approximately two weeks and are usually spread out as evenly as possible for those students who are taking the exams within the regular schedule of the respective major. Nonetheless, students would have multiple exams during a week, sometimes on subsequent days, which would impede the adjustment of their study schedule and hence attenuate my results.

[2] As the experimenter, I was neither involved in teaching nor in grading the courses. Students were informed multiple times that I was providing this service as an external person and that they could not infer anything about the exam content from my practice questions.

[3] Grade reporting was voluntary since the data protection guidelines of the university do not permit the use of personalized administrative data.

surveyed personality traits included patience, risk-aversion, the big 5 personality traits, self-efficacy, and feedback aversion. Furthermore, students stated how motivated they were for their studies in general. Finally, participants had to state how many hours per week they had spent so far studying for this course.

I elicited such a rich set of background characteristics for two main reasons. First, some of the characteristics were ex ante of interest as dimensions of heterogeneity based on results from other studies as well as theoretical considerations. This included gender, ability, socio-economic background, and personality traits. Second, since I could not predict how many students would sign up for the experiment, I wanted to make sure that the design itself would help to increase statistical power. To this end, I asked for the characteristics that should theoretically predict the outcomes well such that I would then be more likely to detect a change in outcomes caused by the treatment.

Furthermore, all outcomes were elicited during this first questionnaire. This again was intended to help with statistical power in the later analyses since initial levels of the outcomes are most likely very good predictors of the final outcomes and can hence help to improve the precision of the estimates.

One main outcome is motivation to study for the exam. It was elicited using a survey question with the following wording: 'How motivated are you to study for the exam of [name of course]?'. Students could choose one value from a 5-point scale: 1 'not motivated at all', 2 'rather not motivated', 3 'neutral', 4 'rather motivated', and 5 'very motivated'. The same wording and scale was used every time this outcome was elicited.

In addition, students were asked about their study plan. This first elicited how many hours they planned to dedicate to studying for this course on each of the six remaining days before the exam. Furthermore, students were presented with the full list of course chapters as outlined on the course syllabus. From these, they had to choose exactly three topics they wanted to focus on in their remaining study time. Due to the design of the feedback, these could later be connected with the feedback topics (see detailed explanation in section 4.2.3).

Lastly, students reported which performance they expected for the exam as well as the upcoming practice questions. Each assessment of students' performance beliefs contained two elements: absolute and relative expected performance. The absolute performance belief was elicited for exam grades on the German scale (from 1.0 to 5.0) and for the points obtained in the first set of practice questions (out of 20). For the expected relative performance, students had to provide probabilities adding up to 100 for being placed in each quartile of the later performance distribution. For each expected absolute performance, I also elicited a self-evaluation, i.e. an assessment of how good or poor students would consider this performance on a 5-point scale, as in Exley and Kessler (2022): 1 'very poor', 2 'poor', 3 'neutral', 4 'good', and 5 'very good'.

After this first block of the questionnaire, I elicited pre-treatment performance via the practice questions that students had to answer to receive feedback. Participants were presented with five open questions about exam-relevant content. These were adapted to the pandemic situation with online exams and allowed for the same conditions as the later exam (e.g. open book and formula sheets). Students had 30 minutes to answer the practice questions, a visible timer ensured that students would stick to this time frame. All questions were presented on the same screen and contained a brief description of the task itself and an open field where students could type in their answer. Once students had submitted their answers or time had run out, they were forwarded to a last screen that elicited beliefs about their performance in the practice questions they had just answered (again in points out of 20). This measure will be used as a pre-treatment performance belief since it is a much more informed assessment than the one provided prior to working on the questions. For motivation and the study plan, pre-treatment outcomes will always refer to the answers provided before answering the practice questions.

The practice questions were then graded by a group of research assistants whom I provided with sample solutions and instructions for grading. A general score (out of 20) as well as an overall ranking of students was calculated. Furthermore, since each of the practice questions referred to a specific chapter of the respective course outline, students also received a sub-score for each topic. According to these sub-scores, a within-student ranking of topics was created, resulting in an assessment of which topic students had answered best as well as worst. These topics would then be used for the positive and negative feedback element respectively.[4]

4.2.3 Feedback and Second Set of Practice Questions

The core part of the experiment was the feedback treatment that followed four days after the practice questions, i.e. three days before the respective exam. Students again received a (personalized) link to visit a page with an individualized structure (see table 4.1). Students were randomized into three groups: a control group that did not receive any feedback at this point and the two treatment groups with positive and negative feedback in varying order. Randomization was performed on an individual level (by course) after receiving the answers to the first set of practice questions. Also, randomization was stratified by gender since this was the main dimension of interest in terms of heterogeneity.

[4] In case of equality of scores for the best and/or the worst topic, one of them was chosen randomly, depending on whether students put in any text at all. This affected 65 students, i.e. 28.9%, for the best topic and 118 students for the worst topic (52.4%), which amounts to a total of 166 students affected (73.8%). Only 67 students had 2 or more best or worst topics (29.8%), 19 students had at least three best or worst topics (8.4%). For students who left all questions blank, it happened that the best and the worst dimension coincided. In this case, randomization of best and worst topic for feedback was repeated until the topics were different. The latter only applied to two students of whom one did not participate in the feedback round. Excluding the other student does not change any of the results. The procedures were applied to students from both treatment groups and the control group.

Participants from the control group did not receive any feedback on the practice questions performed four days before. These students were immediately directed to questions about performance beliefs (including a self-evaluation) and motivation before moving on to the second set of practice questions. Since the main focus of this study was the effect of feedback ordering on the respective outcome, the control group was undersampled and comprised of 20% of the sample. Its main purpose was to observe time trends such as study progress that were independent of feedback as well as to replicate results from the literature showing the effect of feedback compared to no feedback.

The remaining participants were equally distributed between the treatment groups *positive-negative (POSNEG)* and *negative-positive (NEGPOS)*. These students first saw their personal feedback before moving on to the second set of practice questions. Students were informed that their performance on the first set of practice questions had been corrected and evaluated. Then they were presented with the first feedback element which was either positive or negative depending on their treatment group. According to the individual within-person ranking described in section 4.2.2, the positive (negative) feedback element would inform students about the topic they had performed best (worst) in. The respective feedback was communicated in the following way: 'According to all topics evaluated on your performance in the first set of practice questions, you did best/worst on XXX', where XXX stood for their personal best/worst out of the respective topics. For the positive part, an additional sentence was added: 'This is your personal strength, great!'. Similarly, for negative feedback: 'This is your personal weakness, what a pity!'. Furthermore, a green (red) thumbs up (down) was added to make the positive and negative nature of feedback more salient. Examples of how positive and negative feedback elements looked like can be seen in figures A4.1 and A4.2 in the appendix (translated from German). The positive and negative feedback parts were split onto separate screens and would be seen at a distance of a few minutes.

Between and after the two pieces of feedback, participants from the two treatment groups were asked about all relevant outcomes considered in the later analyses. This included their current motivation to study for the exam assessed by the same survey question as in the first questionnaire, their performance beliefs as elicited four days before, and their study plan (hours and topics). Additionally, students were asked about their feelings regarding the single feedback elements. More specifically, they were asked after each feedback element: 'How do you feel about the feedback you just received?' and had the answering options 'very bad' (1), 'bad', 'neutral', 'good', and 'very good' (5).

After receiving feedback and responding to questions on all outcomes, all students answered an additional set of practice questions. The provision of such a second set of practice questions had the goal of being able to measure any potential treatment effects on immediate performance. This is an interesting outcome in itself and also stands in contrast to the exam a few days later: students didn't have time to prepare for these questions after receiving feedback, but they were also associated with relatively low stakes. The questions consisted of

10 multiple choice questions related to the exam content.[5] Each question had a time limit of 60 seconds in order to be more similar to the later exam situation under time pressure. Students could choose exactly one out of three options for each question and would receive one point for each correct answer. No penalties for wrong answers were applied. Each multiple choice question was presented on a separate screen that included a timer indicating the remaining answer time.

Lastly, all students were automatically informed about their score in the second set of multiple choice questions. After responding to the last question, participants were forwarded to a screen thanking them for their participation and informing them about their total number of correct answers out of 10.[6] Hence, students from the control group in the end also received some feedback but only on their total number of points in the second set of multiple choice practice questions. This implies that comparisons of any outcome measured after this point in time between the treatment groups and the control group measure the effect of additional, more detailed feedback for the treatment groups.

4.2.4 Post-exam Questionnaire

One day after the exam, students received the link to a post-exam survey. It contained questions about their exam performance beliefs (including a self-evaluation), the study plan they had actually implemented (hours dedicated to studying for this exam on each of the six days before the exam and topics on which they focused), and their perception of the difficulty of the practice questions provided in the experiment compared to the exam. It also asked about the perceived usefulness of the practice questions as an exam preparation, a recall of the feedback, questions about how they felt about the feedback in general, and how useful they thought the feedback was.

These outcomes were of interest by themselves, but also had the aim of uncovering potential mechanisms of treatment effects. Whether students remember the feedback and its ordering might play an important role for how immediate effects could translate into medium-term effects related to the exam. Furthermore, students' feelings about the feedback in the longer run might contain an emotional component of feedback that would be expected from psychological theories. Similarly, if students perceived the practice questions or the feedback as not very useful, this might explain their reactions to the feedback. Finally, their actually implemented study plan gives a glimpse into their behavior between the feedback intervention and the exam. This might be an important explanatory factor for any effects on exam grades.

[5] In one of the two courses, there were only multiple choice questions in the later exam, even though not as time-constrained. The later exam of the other course consisted mostly of questions as presented in the first set of practice questions. This information could be inferred from previous exams held while teaching online during the pandemic and was hence known when designing the treatment.

[6] In fact, a large majority of participants still remembers their score in these multiple choice questions after the exam as will be discussed in section 4.4.5.

As soon as official grades were released, students were contacted again. They were asked to report their grade in the respective course together with some proof, e.g. a screenshot of their transcript. Each student who reported her grade then received a 10 Euro voucher after I had checked the grade proof.

Exam grades are an interesting outcome for a series of reasons. First, they are a high-stake outcome for the students. This is especially true in this setting where students are at a very early stage of their studies and hence receive one of the first external assessments of their suitability for their university studies. Considering that this took place during the pandemic where students had only experienced online teaching, additional feedback might have been especially valuable to them and hence may have had particularly strong effects. Second, compared to the immediate performance in the multiple choice questions answered right after the feedback, the exam provides a medium-term learning outcome. Students had another three days after the feedback to study for the exam. In these remaining days, they might have adjusted their study plan on two margins: quantity and quality. Adjusting the quantity would imply increasing or decreasing the study hours while the quality or efficiency of study could be reflected in the content they focus on while studying.

4.3 Sample and Descriptive Statistics

This section gives an overview of the participants of the study, including their descriptive statistics and balancing checks for background characteristics and pre-treatment outcomes.

Table 4.2 shows the number of participants and the corresponding attrition rates between the different stages of the experiment, by course. The main sample for the analysis consists of 225 students who participated in the feedback intervention (see column 5).[7] Although attrition rates were rather low from one stage to another, the final estimation sample reduces to around 63% of those who initially registered for the experiment. The numbers about grade reporting in column 7 are related to the main sample in column 5 since grade reporting did not depend on answering the post-exam questionnaire and was incentivized separately.

Table 4.3 shows attrition rates by treatment status. The sample after the first questionnaire comprised of 266 students that were randomized into one control and two treatment groups as described above, by course and gender. The control group consisted of 20% of the sample, i.e. 51 students. Of those, 44 students actually received the feedback in the second stage. Some students had participated in the experiment for both courses such that for them only the observation from course I is used. The remaining 42 non-duplicate observations are then part of the analysis sample.

[7] These 225 students consisted of 86 participants from course I and 139 students from course II, 145 minus the duplicate observations from 6 students who participated in the experiment for both courses. For these 6 students, only the observation from course I was used since it had all stages exactly one day before course II.

The *positive-negative* (POSNEG) and *negative-positive* (NEGPOS) treatment groups comprised of 40% of the sample each, i.e. 106 students for POSNEG and 109 for NEGPOS. Of these students assigned to the treatment groups, 89 (92) finally participated in the treatment for the POSNEG (NEGPOS) group, excluding one of the observations for the duplicate students who participated in both courses. Participation in the second questionnaire (column 2) was not related to the later treatment status of students (not shown). Similarly, attrition in the post-exam questionnaire (column 4) is not related to the realized treatment status (not shown).

Incentivized grade reporting was offered to everyone from the main sample, i.e. those who participated in the second set of practice questions (column 3). Hence, all relative numbers in column 5 refer to this group. Although a slightly larger attrition rate can be observed for the control group, the difference to any of the treatment groups separately or jointly is not statistically significant (not shown).

Table 4.4 shows descriptive statistics for the main control variables to be used in some specifications of the regression analyses. The final estimation sample is almost gender-balanced and has a median and mean age of 20 years. This is in line with students being recruited from two courses that are meant to be taken in the first and in the second year of their undergraduate studies. Participants mostly come from Business Administration and Economics majors, a small part is enrolled in Business and Economics as a minor or other majors (mostly teaching degrees for economics). Most students are in the semester corresponding to the study plan for their major, i.e. at a very early stage of their undergraduate studies. 90% of the sample obtained their high school degree in Germany and hence provided a German high school GPA and their last math grade from high school. Grades are presented on the German scale (1, best, to 6, worst) and mirror the numerus clausus that the university applies when selecting candidates. 55% (63%) of students state that their mother or other female guardian (father or other male guardian) has a university degree.

This points at a rather more academic sample in the experiment compared to Business and Economics students in general as can be found in the 5th cohort of the NEPS university students data (see NEPS-Netzwerk (2021)). There, 31% of first-semester Bachelor and Economics students have at least one parent with academic background (71% in my sample). This might also reflect the setting in the city of Munich where living costs are comparably high and hence the student population might be highly selected to begin with. Contrarily, my sample also has a rather high level of migration background: in the NEPS data, 18% of Business and Economics students are born abroad or had a parent who was born abroad, while these students make up 36% of my sample.

Furthermore, students self-reported their patience, risk-aversion, big 5 personality traits, feedback aversion and self-efficacy, on 5-point scales. Students consider themselves rather patient and risk-averse and they assign quite high scores for conscientiousness, openness, extraversion, and self-efficacy. Instead, they state to be less neurotic and agreeable. Furthermore, participating students attribute themselves a low level of feedback aversion.

Finally, students were asked how motivated they were to study for their university studies in general as well as how many hours they had spent per week studying for the specific course so far. Motivation was measured from 'not motivated at all' (1) to 'very motivated' (5) and students on average seem to be quite motivated. Also, students on average invested 6.5 hours per week to study for the respective course (with a median of 5 hours and one large outlier at 135).

A similar descriptive table for pre-treatment outcomes can be found in table A4.1. Self-reported motivation to study for the exam again was measured via a survey question that asked students how motivated they currently were to study for the respective exam. The scale ranged from 1 ('not motivated at all') to 5 ('very motivated'). Overall motivation was relatively high with a mean of 3.5 and a median value of 4.

Students' pre-treatment performance was measured using their score in the first set of practice questions. Out of 20 possible points, students on average scored 8.5, implying that students were still lacking quite some course knowledge one week before the exam. Interestingly, students had higher expectations regarding their points before answering the questions (12.6 points on average), but adjusted these expectations downwards after having responded to the practice questions (mean of 8.7 points). This immediate absolute performance belief is on average quite accurate. The self-evaluation of this performance was assessed by asking students whether they would consider this expected performance 'very poor' (1), 'poor', 'neutral', 'good', or 'very good' (5). Again, students considered their expected performance much better before answering the practice questions (mean of 2.9 before compared to a mean of 1.9 after).

Expectations about future exam grades were also elicited before treatment, i.e. seven days before the exam. The German scale for grades ranges from 1 (best) to 5 (fail) and only contains decimal values at .3 and .7 between integer values. Interestingly, no student expected to fail the class one week before the exam. On average, students expected a grade of 2.3 which would be classified as 'good'. Self-evaluation of exam grades follows more closely the self evaluation of the first set of practice questions before answering them and is even more optimistic with a mean of 3.2.

Finally, students were also asked to state their study plan during the first questionnaire. On the one hand, this included questions on their planned study hours on each of the remaining days before the exam. On the other hand, students were asked to pick three topics from the course syllabus that they were planning to focus on in their remaining study time. Students planned to steadily increase their study hours when moving closer to the exam with the highest value on the last day before the exam (4.2 hours on average). Furthermore, 27% (29%) of students had the topic they would perform best (worst) in on the later practice questions on their list of prioritized study topics.

Table 4.5 presents a balancing check of all controls for the three groups students were randomized into. Columns 4 to 6 show differences across the treatment and control groups, including stars indicating a significant difference. As we can see, only three differences turn out to be significant at the five percent level which could as well be expected by chance. Similarly, a joint F-test on all differences cannot reject the hypothesis that they are jointly zero (not shown). Furthermore, all variables seen in this table are controlled for in some specifications of the presented regressions.

A similar balancing check for pre-treatment outcomes can be found in table A4.2. Pre-treatment motivation and performance did not significantly vary between the control and treatment groups. Some minor differences can be observed for performance beliefs about the first set of practice questions. Students in the POSNEG group overall seem to be slightly more optimistic before answering the questions which is less visible after all students have actually done the practice. Both treatment groups are mildly more optimistic about their future grade (on a German scale) compared to the control group. No differences can be observed regarding students pre-treatment study plan. In the later regressions, I will mostly control for the respective pre-treatment outcome which should take care of any remaining imbalance.

4.4 Results

This study aims to evaluate the effects of feedback ordering on the main outcomes motivation and performance as well as the further outcomes beliefs and study behavior. I will first show results on motivation (section 4.4.1), then move on to beliefs (section 4.4.2) and the study plan (section 4.4.3) before showing the results on performance (section 4.4.4). Section 4.4.5 explores some further mechanisms of the treatment effects, whereas section 4.4.6 summarizes heterogeneous treatment effects that are described more in detail in appendix A4.1. Finally, section 4.4.7 looks at the effects of receiving any feedback.

4.4.1 Treatment Effects on Motivation

Students were asked multiple times how motivated they were to study for the exam of the specific course for which they were participating in the experiment. All students received this question before solving the first set of practice questions (seven days before the exam), control group students were asked again once before solving the second set of practice questions (three days before the exam). Students from the two treatment groups received the question about topic-specific study motivation twice before solving the second set of practice questions (three days before the exam), after each of the separate feedback elements (see table 4.1). This setup allows to estimate various types of effects: the effect of feedback ordering by comparing post-feedback outcomes for the two treatment groups POSNEG and NEGPOS, the effect of positive versus negative feedback only with a between-feedback comparison of

the two treatment groups, and the effect of the positive and negative feedback respectively compared to no feedback by comparing treatment groups to the control group.[8]

I start by looking at the effects of feedback ordering on motivation across treatment groups. To do so, I use the following basic regression that only includes the two treatment groups who received feedback on the first set of practice questions:

$$O_i = \beta_0 + \beta_1 POSNEG_i + female \times course + best_i + worst_i + \beta_2 O_{i0} + \beta_3 performance_{i0} + \mathbf{X}'_i \boldsymbol{\beta}_4 + \varepsilon_i \tag{4.1}$$

O_i refers to the different outcomes as described above, in this case motivation. $POSNEG_i$ is a dummy that takes the value one for the treatment group with positive feedback first and negative feedback afterwards, and zero for the group receiving negative feedback first. ε_i is an individual error term, I use robust standard errors. Furthermore, equation (4.1) contains dummies for the respective course-gender cell within which I randomized as well as for the respective topics students received feedback on since these were not random in this setting. Also, most specifications will include performance controls from the practice questions as well as the respective pre-treatment outcome.[9]

In different specifications, I then sometimes include a larger set of controls to increase the precision of my estimates. $\mathbf{X}_i$ includes the background characteristics described in section 4.3 (age, gender, semester, field of study, high-school GPA, last math grade in high school, parents' education and occupation, migration status) as well as the personality traits elicited prior to the first set of practice questions (patience, risk-aversion, big 5, self-efficacy, feedback aversion). Not all students reported high-school grades, depending on whether they had completed their high-school degree in Germany. For all those with missing values, I impute high-school performance with the mean of the sample and add an imputation dummy to all respective regressions.

Table 4.6 presents the results of this estimation with motivation as an outcome. Columns 1-3 show the results for motivation measured between the two feedback elements, columns 4-6 refer to the question on motivation after both feedback elements have been provided. All regressions only include students from the two treatment groups POSNEG and NEGPOS. As indicated by the coefficient for the POSNEG dummy, receiving feedback in the ordering *positive-negative* compared to *negative-positive* leads to higher motivation to study for the exam. The effect emerges after the first feedback element has been presented, as can be

[8] All three types of analyses (for all presented outcomes) were pre-specified in the pre-analysis plan that is part of the AEA RCT Registry entry AEARCTR-0008953.

[9] Results look very similar when simply controlling for the female and the course dummies individually, without adding the interaction term (not shown). Similarly, results do not change much if I include the respective points in the best and worst topic as a pre-treatment indicator of performance rather than overall performance in the first set of practice questions (not shown).

seen in columns 1-3. This can be interpreted as the pure effect of positive versus negative feedback as single elements, which implies that the information content of the feedback is quite different.

After students have received the second feedback element as well, this information asymmetry is no longer there since students from both groups were informed about the best and worst topic of their performance in the practice questions. Hence, the estimates in columns 4-6 show the treatment effect of the feedback ordering on study motivation. The effect is positive and significant, which is especially evident when including pre-treatment motivation and performance. The latter are strong predictors for post-treatment motivation as can be seen from the values of the R-squared in columns 2 and 5 and hence they increase statistical power of the estimation.

The measure for motivation is standardized here such that the treatment effect can be interpreted in standard deviations. This implies that first receiving positive feedback leads to almost 0.3 standard deviations higher post-feedback motivation to study for the respective exam compared to first receiving negative feedback in my preferred specification in column 5. Including a large set of controls in columns 3 and 6 increases the coefficient size, but also loses two observations due to missing controls.[10]

Overall, coefficient sizes vary to a certain degree across columns 3-6 of table 4.6 (and in the full table A4.5) which might raise concerns about the validity of the results. The changes might be related to the relatively low sample size: even if carried out correctly, randomization in small samples may still produce some correlation between the treatment and the control variables. This could explain why coefficient estimates are affected by the inclusion of controls. Including many controls should then in principle reduce omitted variable bias when estimating the treatment effects. In this rather small sample though, the inclusion of many controls may lead to overfitting of the specification. If that were an issue here, we would see the standard errors of the coefficient on the POSNEG dummy increase which does not seem to be the case in column 6 of table 4.6.

To exclude that the inclusion or exclusion of certain controls drives the observed treatment effects, I perform both LASSO and double LASSO procedures that choose covariates to be included in the model based on their relevance for predicting the outcome only (LASSO) or both the outcome and the treatment variable (double LASSO). Table A4.3 shows the treatment coefficient of the resulting estimations with selected covariates. For both methods, I run two variants: one forces the procedure to include the main indicators from column 5 of table 4.6 (female $\times$ course indicator, feedback topic dummies, pre-treatment motivation and performance) into the estimation, the other one allows for free choice among all indicators, dummies, and controls. The coefficient size of the preferred restricted versions in columns 1 and 3 in table A4.3 is very close to column 5 in table 4.6 which confirms choosing this as

[10] Tables A4.4 and A4.5 show a full set of specifications for both outcomes, including a raw comparison without any controls. It then adds all respective controls separately.

a preferred specification. When allowing the selection procedure to choose freely from all controls, coefficients on the treatment dummy become smaller and insignificant, standard errors change very little though.

To understand what is driving the treatment effects on motivation, I next compare the evolution of motivation in the treatment groups to the initial motivation of the control group. Borrowing from Coffman et al. (2021), I estimate the following equation:

$$\begin{aligned} O_{it} = \beta_0 &+ \beta_1 POSNEG\_initial_{it} + \beta_2 POSNEG\_fb1_{it} + \beta_3 POSNEG\_fb2_{it} \\ &+ \beta_4 NEGPOS\_initial_{it} + \beta_5 NEGPOS\_fb1_{it} + \beta_6 NEGPOS\_fb2_{it} \\ &+ \beta_7 CONTROL\_after_{it} + female \times course + best_i + worst_i + \mathbf{X'}_i \boldsymbol{\beta}_8 + \varepsilon_{it} \end{aligned} \tag{4.2}$$

All outcomes are standardized with respect to the initial average outcome of the control group which serves as a reference category, i.e. the respective outcomes are demeaned and divided by the standard deviation of the pre-treatment control group outcome. O_{it} refers to the respective outcome (here motivation) at three points in time: pre-treatment (initial), after the first feedback part (fb1) and after the second feedback part (fb2). The coefficients β_1 to β_7 refer to dummies that indicate whether an observation belongs to one of the three groups and when it was measured. This way, the resulting coefficients report deviations from the initial control mean of the respective outcome. The control variables are equivalent to equation (4.1), standard errors are clustered at the individual level.

Figure 4.2 shows the resulting coefficients for the two treatment groups as well as the control group. First, it can be seen that there is no significant difference between the pre-treatment motivation of the three groups of students. Furthermore, there is no movement over time of the control group's motivation (gray plots). Instead, for the NEGPOS treatment group there is a large drop in motivation following the first, i.e. negative, feedback. This does not fully recover to the initial levels of the control group after the second, i.e. positive, feedback, and remains statistically distinguishable from these initial levels. More interestingly though, this drop in motivation cannot be observed for the POSNEG group when they receive the negative feedback part (fb2). This suggests that first receiving positive feedback shields respondents from experiencing a decrease in motivation after negative feedback. Maybe surprisingly, motivation does not increase (much) after receiving positive feedback for the POSNEG group compared to the control group's initial motivation. This might hint towards a ceiling effect: students were initially very motivated to study for the exam and might hence not react much to positive feedback on this dimension.

The results for motivation have some important implications. Positive and negative feedback individually have very different effects on students' motivation in this setting. Positive feedback doesn't affect motivation while (early) negative feedback reduces motivation. Often though, feedback givers do not have the choice between either positive or negative feedback,

but they rather want to provide suggestions for improvement that necessarily imply some negative feedback. My findings suggest that there is no need to avoid giving negative feedback when both negative and positive feedback are provided. Simply paying attention to the ordering of the feedback elements might be a solution to avoid demotivating students with negative feedback.

4.4.2 Treatment Effects on Beliefs

To understand whether the observed effects of feedback ordering on motivation can be explained by belief updating, I next turn to the results on performance beliefs. Performance feedback is often associated with updating beliefs about one's performance, which in turn might have an impact on motivation. Beliefs about performance were elicited on three measures: performance in the first set of practice questions, in the second set of practice questions, and in the exam. All belief elicitation blocks contained three elements: students had to first state which exact points/grade they thought they had achieved/would achieve. Secondly, they had to evaluate this expected performance as either poor or good as described above. Lastly, students had to provide probabilities adding up to 100 about the likelihood of being ranked in each quarter of the respective performance distribution of all participants.

Performance beliefs were surveyed both between the two feedback elements as well as after the second feedback. Since feedback was given on the first set of practice questions, between-feedback belief elicitation focused on the performance in the first set of practice questions and in the exam. Instead, beliefs about performance in the second set of practice questions were elicited only once, right before performing the questions. This leads to three belief outcomes to study: retrospective beliefs about performance in the dimension feedback is given on (right after performing the questions as well as after each of the feedbacks), immediate forward-looking beliefs about the second set of practice questions, and medium-term forward looking beliefs about the exam grade. In the following, I will present results on absolute performance beliefs only.

Table 4.7 shows the results for all explicit beliefs about performance regarding the points achieved in both of the practice sets provided during the experiment as well as the exam grade. It uses equation (4.1) for my preferred specification from col. 5 of table 4.6 which includes a female $\times$ course indicator, feedback topic dummies, and pre-treatment expectations and performance, but no further controls. Both forward-looking beliefs about future exam performance as well as retrospective beliefs about past performance in the first set of practice questions are higher in the POSNEG group compared to the NEGPOS group after the first feedback element, although more convincingly for beliefs about practice set 1. The effect does not persist after the second feedback element for both practice set 1 (on which feedback was given) as well as exam grades. No effect can be seen for the second set of practice questions where beliefs were elicited only once right before answering the questions which was right after the two feedback elements. This implies that incorrect belief updating cannot explain

the treatment effects on motivation since there is no treatment effect of feedback ordering on performance beliefs.

Figure A4.3 shows the evolution of performance beliefs about the first set of practice questions compared to the initial mean of the control group. Interestingly and in contrast to the reactions to motivation, when updating beliefs students follow patterns known from motivated beliefs: they react strongly to the positive feedback, but much less to the negative feedback. This is especially true for students receiving the positive feedback as a first element whereas students who first receive negative feedback react more cautiously to positive feedback, even though the difference between the reactions of the two treatment groups to positive feedback is not statistically significant (not shown).[11]

The results on performance beliefs provide some important insights to the mechanisms behind the treatment effects on motivation. Belief updating has been extensively studied in economic theory such that there is a clear idea on how students would react to the feedback elements I provide. The fact that students seem to react as expected, helps to disentangle mechanisms of the treatment effect on motivation that can be explained by economic theory from other potential mechanisms that we currently don't have good explanations for. Such other mechanisms could be found in other disciplines such as psychology where theories suggest emotional reactions to feedback. These will be analyzed in further detail in section 4.4.5 using data from the post-exam questionnaire.

4.4.3 Treatment Effects on Study Behavior

To better understand potential mechanisms of treatment effects on *performance*, participants were also asked about their study plan. This included questions about the hours they were planning to study for the respective exam on each of the remaining days before the exam. These were also surveyed after the exam to see the actually implemented study behavior of students. Furthermore, students were asked multiple times to name three topics from the course that they were planning to focus on in the remaining study time. These could be chosen from the list of chapters included in the respective course. Participants from the treatment groups received these questions three times: before feedback, after feedback1 and after feedback2. Control students only responded twice to these question blocks as for motivation and beliefs, once seven days before the exam and once three days before the exam.

Table A4.8 in the appendix shows the treatment effect on the aggregated study hours students planned to spend on this course in the last two days before the exam. The coefficients are not

[11] To test formally whether the reactions to positive feedback are different across the two treatment groups, I can compare the coefficient on the dummy $POSNEG\_fb1_{it}$, i.e. the reaction of the POSNEG group to the positive feedback, to the isolated reaction of the NEGPOS group to the positive feedback which is given by the difference between $NEGPOS\_fb2_{i,t}$ and $NEGPOS\_fb1_{i,t}$.

statistically distinguishable from zero and if anything negative for those receiving positive feedback first.

Instead, figures 4.3 and 4.4 illustrate how students reacted to the topics they have been given feedback on. For each student, positive and negative feedback was given individually since it depended on which of the topics they had performed in best and worst during the first set of practice questions. The measure is also available for control group students since their practice questions were corrected as well. Hence, they also have a best and worst topic, but didn't receive any feedback on them. This allows me to construct an indicator for each student at each time they responded to this question that is 1 if the student has the respective best/worst topic on their list of top three study topics and 0 if not. Then, I can repeat the estimation from equation (4.2) for this indicator at three (two) points in time for the treatment (control) groups.

Figure 4.3 shows the evolution of this indicator for the topics individuals performed in best and eventually received positive feedback on if they were in one of the treatment groups. At the initial stage, i.e. pre-treatment, there is no significant difference between the control and treatment groups on how likely individuals were to have this individual-specific best topic on their list of study priorities. After receiving positive feedback on the respective topic, the POSNEG group shows a significant drop in the share of individuals who include this topic on their study list compared to the initial control mean. Similarly, this happens for the NEGPOS group after they have received positive feedback, even if not significantly so compared to the initial control mean. This suggests that students are more cautious to remove a topic from the study list if they received negative feedback before.[12] No change can be seen for the control group, confirming that the results for the treatment groups do not include a general trend.

Interestingly, this development is still visible when looking at students' responses to the question which topics they had actually focused a few days later (see figure A4.4 in the appendix). Even one day after the exam, individuals report less often to have this personally best topic as part of their study list in the POSNEG group, but not significantly so in the NEGPOS group. The coefficients for both groups are smaller in magnitude and significantly different from the coefficients after the second feedback element, at the 5% level for the POSNEG group and at the 10% level for the NEGPOS group. This suggests that students face some constraints in following their updated study plan, but it might as well mean that students remember their positive feedback topic less accurately. Again, no general trend can be seen for the control group.

Similarly, figure 4.4 shows how the inclusion of worst topics onto individuals' study lists evolves over the course of the experiment. Again, students react to the feedback provided during the experiment: in this case, they add topics to their study list they received neg-

[12] In fact, the difference between $POSNEG\_fb1_{i,t}$, i.e. the reaction of the POSNEG group to the positive feedback, and the isolated reaction of the NEGPOS group to the positive feedback, $NEGPOS\_fb2_{i,t} - NEGPOS\_fb1_{i,t}$, is significantly different from zero.

ative feedback on. This seems to be equally strong for both treatment groups, the coefficients for $NEGPOS\_fb1_{i,t}$ and the difference of the coefficients for $POSNEG\_fb2_{i,t}$ and $POSNEG\_fb1_{i,t}$ are not significantly different from each other. The control group again does not show any changes which is plausible given that they never received the information on their best and worst topic.

Similarly to the positive feedback topics, the results for negative feedback topics persist to the post-exam questionnaire about actual study topics as well (see figure A4.5). Again, effect sizes become smaller and are significantly different to the coefficients from before the exam. But they are still significantly different from the initial control mean and there is no development for the control group.

The results for students' study behavior again help to get a better understanding of the mechanisms during the experiment. The fact that students do not adjust their study time suggests that there is limited flexibility in students' learning schedule, but might also imply that they consider other margins of adjustment more relevant in this context. In fact, students seem to rather adapt their study efficiency than the quantity by changing the topics according to the feedback they received. This has important implications for highly time-constrained settings in which students might still benefit from feedback although they cannot increase their study time. Since students react in a similar fashion in both treatment groups, this finding can be generalized to receiving any feedback rather than being unique to a specific feedback ordering.

4.4.4 Treatment Effects on Performance

The treatment effect on performance can be analyzed using two measures: immediate performance in the second set of practice questions (points out of 10) and medium-term high-stake performance in the exam (German-scale grade). Both measures presented here are standardized across treatment groups to have a mean of 0 and a standard deviation of 1. This way, coefficients can be interpreted as changes in standard deviations. Furthermore, exam grades are inverted such that higher grades correspond to better performances. This is necessary because in the German grading scale smaller numbers correspond to better grades.

Table 4.8 shows the treatment effects on both performance measures according to an estimation of equation (4.1). Columns 1-3 focus on immediate performance in the second set of multiple choice practice questions. Coefficients are statistically not distinguishable from zero and if anything negative. When turning to the grades in the respective exam (col.s 4-6), a similar picture emerges: coefficients are now slightly positive, but still insignificant. From simply looking at the confidence intervals, I can exclude effect sizes larger than 0.358 at the 95% confidence level (0.316 at the 90% level, 0.441 at the 99% level). These results imply that the changes in motivation as well as the adjustment of study topics do not necessarily translate into changes in immediate or medium-term high-stake performance.

A reason for this could be the relatively short time frame. Students do not have the time to prepare further for the second set of practice questions since they answer them right after they have received the feedback. Fischer and Wagner (2018) show in a field experiment that students react negatively to feedback given right before an exam and positively to similar feedback given a few days earlier. These results are in line with the direction of the coefficients, although they are not statistically different from zero.

Another potential reason for not finding treatment effects is the lack of statistical power in detecting relatively small effect sizes. To be able to detect an effect size of 0.099 from my preferred specification in column 5 with a confidence level of 90% and statistical power of 80%, I would need a total sample size of around 2,500 students. This could be reduced to about 800 students when considering the large explanatory power that the control variables of the regression have.[13]

Additionally, the sample size of the students who voluntarily reported their grade is smaller than the full estimation sample for the other treatment effects. This is caused by grade reporting being voluntary rather than me having access to administrative data for all students. Grade reporting was incentivized with a 10 Euro voucher and around 75% of students sent their grade, including a proof from their student portal (see table 4.2).

This raises concerns about sample selection into grade reporting. To get an idea of such potential selection, I compare the grades reported by students in the experiment with the aggregated statistics of all students that can be found in the publicly released grade distribution.[14]

Figure 4.5 shows the experimental and the official cumulative distribution functions for all students who passed course I. Grades on the x-axis are on a German scale, i.e. from 1.0, best, to 4.0, worst. The students who reported their grades in my experimental sample seem to be slightly positively selected in all parts of the distribution. The median grade is 3.3 in both distributions, but the share of students with the lowest grade (4.0) is higher in the full student population (around 30%) than in the experimental sample (around 20%). The share of students who failed the course differs markedly between the two groups: only 22.7% of students in my sample provided proof of them failing the course whereas official statistics report that 50.8% failed the course. Since the difference is so stark on this dimension but not in the group of students who passed the exam, this might imply that students who failed the course disproportionately often did not report their grade in the experiment.

[13] The underlying calculations have been performed with the STATA command *power* as well as with Optimal Design Software, see Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org (last accessed on October 28).

[14] Official university statistics only report the graphs in panels B of figures 4.5 and 4.6, without the underlying data, and, separately, the respective passing rates. I will hence consider selection separately by course using the available plots.

To shed more light on this argument, table A4.20 shows a balancing check for students in course I, distinguishing by whether students reported their grade or not after having participated in the treatment. The results are rather surprising since it seems that students with a lower feedback aversion tended not to report their grades. Similarly, the students who failed to report their grade more commonly have a mother who is working and has a university degree. Also, they are more likely to have a second-generation migration background and grade reporting does not depend on initial motivation. Less surprisingly, those students who performed worse on the first set of practice questions are less likely to report their final exam grade. Given the amount of comparisons presented here and the often marginal level of statistical significance, these differences could also be observed by chance.

Figure 4.6 shows the distributions for course 2. Here, the positive selection into grade reporting is even more evident: median grades differ by one full grade point (2.0 in my sample vs. 3.0 in the full course) and the share of students who reported the lowest passing grade is around 4% in my sample vs more than 10% in the entire course. The difference between the two distributions is especially visible in the upper half where in my sample 40% of the students perform better or equal than a 2.0 whereas this share is only 20% in the full course distribution. When looking at passing rates, again, students in my sample are more likely to have passed the exam (around 90%) than in the full course distribution (around 58%).

Table A4.21 compares all characteristics used in the regressions between those who did or did not report their grade after receiving treatment in course 2. Students who do not report their grade are more risk-averse and are more likely to have a migration background, this time in the first generation. Furthermore, they have different majors, are less conscientious, and show higher levels of extraversion. Also, they seem to report higher, i.e. worse, high school GPAs which is in favor of the explanation posed above that the lowest performers, i.e. those most likely to fail the exam might avoid reporting their grades in the framework of the experiment, even though they are paid for it. Again, grade reporting does not depend on initial motivation but is more likely for students with higher performance in the first set of practice questions.

Generally, the evidence from both courses is consistent with a notion that the money I offered to students for reporting their grades (10 Euro vouchers for Amazon or Avocadostore) did not fully compensate for other reasons why students might want to avoid reporting their grades. In particular, students with higher risk aversion might have been more cautious about potential consequences or might feel shame associated with very bad outcomes. Similarly, students with relatively highly educated mothers, as in course I, might be less willing to communicate a very bad grade. This could e.g. be driven by the fact that the monetary reward is less relevant for them or come from a higher feeling of failure when receiving a worse grade if higher education in the household is associated with more pressure about exam grades. Interestingly, no differences by gender can be observed.

Additionally, the reaction of students regarding their performance in the exam might depend on which of the (feedback) topics were included in the final exam. In table 4.9, I hence show

an exploratory analysis using the course topics present in the final exam.[15] The results from columns 4-6 of table 4.9 show that there is a positive treatment effect of the POSNEG order compared to NEGPOS for those students who faced their worst practice topic in the exam. This applies to 67/140 students, i.e. half of the estimation sample, and is quite evenly distributed among the treatment groups (not shown). In general, not having the worst feedback topic in the exam seems to benefit students. Furthermore, not having the worst topic in the exam reduces the benefit of the POSNEG order. No clear pattern can be observed for the positive feedback topics (col.s 1-3).

These results imply that the effect of feedback ordering on exam grades might depend on which topics students actually face during the exam. Students from both treatment groups were equally likely to add the negative feedback topic to their study list after hearing they had performed worst in it. If this topic then does not show up in the exam, students might feel differently about the adjustment of their study content. This might in turn affect their exam performance if focusing on this topic does not pay off for the exam. The latter might be especially true for students who received the negative feedback as a second element and hence were more motivated to study for the exam than those from the NEGPOS group. If these POSNEG students were especially eager to follow their adjusted study plan, they might be relatively more frustrated when not finding the worst feedback topic in the exam.

Students who instead encounter the worst feedback topic in the exam, might also perform differently based on the feedback ordering they were assigned to. Participants from the POS-NEG group might remember that their motivation did not suffer from the negative feedback associated with that topic since they received positive feedback before. This might reduce the negative impact of encountering this topic in the exam when heving been in the POSNEG group. Hence, although the effects on motivation do not translate into performance effects on average, they might be a mediator of treatment effects on performance when students are faced with their negative feedback topic in the exam.

4.4.5 Feelings and Thoughts about Feedback

This section looks at students' feelings and thoughts about the feedback I provided, which could uncover important additional mechanisms for the observed treatment effects. The outcome measures come from two parts of the questionnaires. Feelings about the respective feedback elements were elicited right after each feedback part three days before the exam. All other outcomes presented in this section come from the post-exam questionnaire that was sent to students one day after the exam.[16]

[15] I was provided with the full text of the final exams for both participating courses and identified which course topics were part of the exam. From that, I can construct a simple indicator of whether the individually best and worst topics from the practice questions were included in the exam or not.

[16] Around 86% of students who answered this questionnaire responded on the day after the exam, another 5% within two days after the exam. Only around 2% of students answered the post-exam questionnaire 4 or 5 days after exam.

To see whether students fully grasped the presented feedback, I want to present two important checks. The first one refers to the treatment groups only and compares how they stated they felt after each of the separate feedback elements. The scale ranges from 'very bad' (1) to 'rather bad' (2), 'neutral' (3), 'rather good' (4), and 'very good' (5). Table 4.10 shows that feelings after positive feedback (col. 1 row 1 and col. 2 row 2) are generally better than after negative feedback (col. 1 row 2 and col. 2 row 1) for both groups. Furthermore, the differences between the two treatment groups are significant for both feedback elements, i.e. participants generally feel very differently after positive and negative feedback.

Table 4.11 shows some descriptive statistics from the post-exam questionnaire. Students were asked about the difficulty of the exam from 'very difficult' (1) to 'very easy' (5) and about the relative difficulty of the practice questions I and II with respect to the exam ('much more difficult' (1) to 'much easier' (5)). On average, students perceived the exam as rather difficult and the first (second) set of practice questions as similarly difficult (slightly easier) than the exam. Furthermore, on a scale from 'not useful at all' (1) to 'very useful' (5), participants on average rated the first and second set of practice questions as neutral in terms of their usefulness to prepare for the exam. Interestingly, this differs by whether the respective best or worst feedback topic was in the exam. Students whose best feedback topic was part of the exam perceived the first set of practice questions to be much more useful (not shown). Similarly, students whose worst topic was part of the exam perceived the first set of practice questions as significantly less useful. No differences can be observed for the second set of practice questions (not shown).

Furthermore, students were asked to recall the feedback they received on both sets of practice questions. All students received feedback in the form of a final score for the second set of practice questions, but only 89% of students remember or saw this final score. Of these students, 95% correctly recall their final score. Furthermore, these students on average perceive the feedback as 'neutral' in terms of its usefulness.

Finally, students were also asked to recall feedback on the first set of practice questions which only the two treatment groups received. 84% of participants correctly recall having or not having received this feedback. Of those who correctly remember receiving feedback, 88% correctly recall having received both positive and negative feedback. From these remaining 122 participants, 91% remember the ordering of feedback. The usefulness of the feedback on the first set of practice questions was perceived as slightly lower than that on the second set. General feelings about the feedback received on the first set of practice questions were elicited with the question 'How did you overall feel about the feedback you received on the first set of practice questions?'. The scale ranged from 'very bad' (1) to 'very good' (5) and students on average, state to have felt 'neutral'.

Additionally, I look at whether there is an effect of feedback ordering on any of the post-exam outcomes. Table A4.22 shows how all outcomes from the post-exam questionnaire differ by treatment status. Most outcomes don't seem to be affected by students' treatment status.

The exception is the answer to the question regarding overall feelings about the feedback on the first set of practice questions. These overall feelings reported one day after the exam are significantly better in the POSNEG group compared to the NEGPOS group. This is confirmed by a regression of the same specifications used in tables A4.4 and A4.5 with post-exam feelings about feedback as an outcome. Table 4.12 shows the results of these regressions now using standardized post-exam feelings about feedback. For students in the POSNEG group, the measure of overall feelings is almost 0.9 standard deviations higher than for the NEGPOS group in my preferred specification in column 5. This corresponds to 0.7 points better feelings on the 5-point scale. Although these effects are estimated on the slightly smaller sample of students that responded to the post-exam questionnaire, they can most likely be generalized to the main estimation sample since the results from section 4.3 suggest that there was no differential attrition by treatment status. The results on feelings about feedback point to emotions being a potential mechanism for the treatment effects on motivation which is in line with the hypotheses from psychology mentioned above.

4.4.6 Heterogeneous Treatment Effects

In the pre-analysis plan, I specified a series of dimensions of heterogeneity that ex ante were of interest. Due to the relatively small final sample size, the interpretability of these heterogeneous treatment effects is rather limited, such that I report the results in more detail in appendix A4.1. In the following paragraphs, I will summarize some suggestive evidence from these analyses.

A student's gender was the main dimension of interest because of prior evidence highlighting how women react differently to feedback than men (Buser, 2016; Coffman et al., 2021; Goulas and Megalokonomou, 2021). In my experiment, I do not find clear evidence of any gender differences in reacting to feedback ordering for any of the outcomes (section A4.1). Coefficients on the interaction between treatment and the female dummy are mostly positive, with the exception of exam performance where the POSNEG feedback ordering might be harmful for women.

The initial performance level was the second heterogeneity of interest since prior evidence suggests that low- and high-achievers can have different reactions to feedback (Bandiera et al., 2015; Goulas and Megalokonomou, 2021; Hermes et al., 2021). Indeed, I also find that the treatment effect on motivation is smaller the higher an individual's pre-treatment performance was, indicating that relatively lower-achievers are even more motivated by the feedback sequence highlighting their strength first (section A4.1).

A further characteristic that might influence the perception of feedback ordering is a student's socio-economic background (SES). I define SES with a dummy indicating whether at least one parent has university education, hereby assuming that there might be systematic differences in feedback culture between families with and without academic parents. Maybe surprisingly, students from *non-academic* families overall seem to react less to the feedback ordering,

e.g. the treatment effect on motivation is fully concentrated among those with at least one academic parent (section A4.1).

Lastly, individuals' personality traits do not seem to systematically matter for their reaction to the feedback sequence. Some of the coefficients of the respective interaction terms with the treatment dummy are significantly different from zero, but no stable pattern can be observed (section A4.1).

4.4.7 Feedback vs no Feedback

In the following I want to present some evidence on the effect of feedback compared to no feedback. An important caveat to this analysis is the small sample size of the control group that did not receive feedback on the first set of practice questions: it comprises of 44 individuals only. The shares of the respective groups within the final sample were pre-specified and set to 40% of the sample for each of the treatment groups and 20% for the control group. These proportions were set this way to ensure that there would be a sufficient number of observations for the main comparison of the treatment groups with different feedback ordering, at the expense of potential comparisons with a control group not receiving any feedback.

Furthermore, participants in the control group as well as those in the treatment groups received feedback on the second set of practice questions in form of a total of points out of 10. This implies that any observed effect would actually be the effect of the additional feedback for the first set of practice questions compared to just the one on the second set of practice questions. Nevertheless, this comparison is useful in getting an understanding of whether in general the feedback was useful to students and whether the POSNEG (NEGPOS) ordering actually increased (decreased) motivation overall.

Figure 4.7 shows the treatment effects of receiving feedback compared to the control group, both on average across the two treatment groups as well as separately. These values come from the respective regressions of all outcomes either on a dummy for being in one of the treatment groups (black plots) or on dummies for each of the treatment groups separately (blue plots). All regressions contain the full set of controls to increase precision of the estimates. No average effect of receiving feedback vs not receiving feedback on the first set of practice questions is significantly different from zero. Furthermore, none of the coefficients on the separate treatment dummies is significantly different from zero, with the exception of the coefficient on the NEGPOS dummy for motivation (significant at the 10% level). This is in line with figure 4.2 and shows that in this setting, only receiving feedback in the NEGPOS ordering is statistically different from receiving no feedback on the first set of practice questions.

4.5 Conclusion

This field experiment varies the ordering of a negative and a positive feedback element on university students' performance in exam practice questions. I find that first giving positive feedback increases post-feedback motivation to study for the respective exam. Furthermore, students react to the feedback by adjusting their study plan to the content of the feedback elements. I do not find effects of feedback ordering on performance beliefs and average exam performance.

These results hint at feedback operating beyond its pure information channel. Students in both treatment groups get positive and negative feedback on their performance and hence receive the same level of information regarding their pre-treatment level of knowledge. The observed effects therefore suggest an emotional or impulsive reaction, especially to the feedback students receive first.

The present study adds to the literature by looking at dynamic reactions to feedback in a field setting. The exam context of university students as a real-life setting with high stakes can provide important insights beyond a lab setting. This is an important step towards understanding how to give feedback in the education context where learning individuals are especially dependent on their instructors' assessment of their performance.

The results from this experiment can also help to shed light on how to motivate students to pursue their studies. In fact, motivation has shown to be related to self-confidence (Bénabou and Tirole, 2002) and negatively associated with dropout from education (Gillet et al., 2012; Cabus and De Witte, 2016; Rump et al., 2017). My findings show that it is not necessary to cut down on the corrective feedback that aims to help students improve their performance. Rather, to avoid a drop in study motivation, it is important to highlight their personal strengths first rather than afterwards.

Figures and Tables

Figure 4.1 : Overview of the Study Design

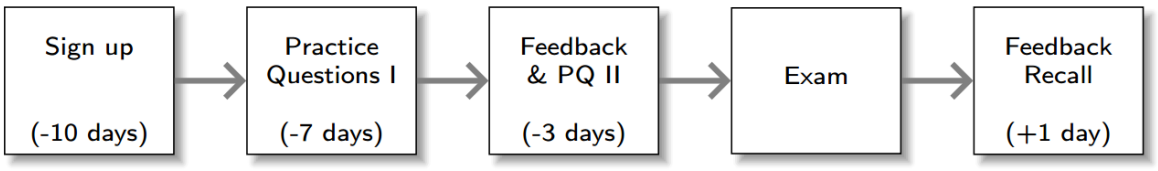
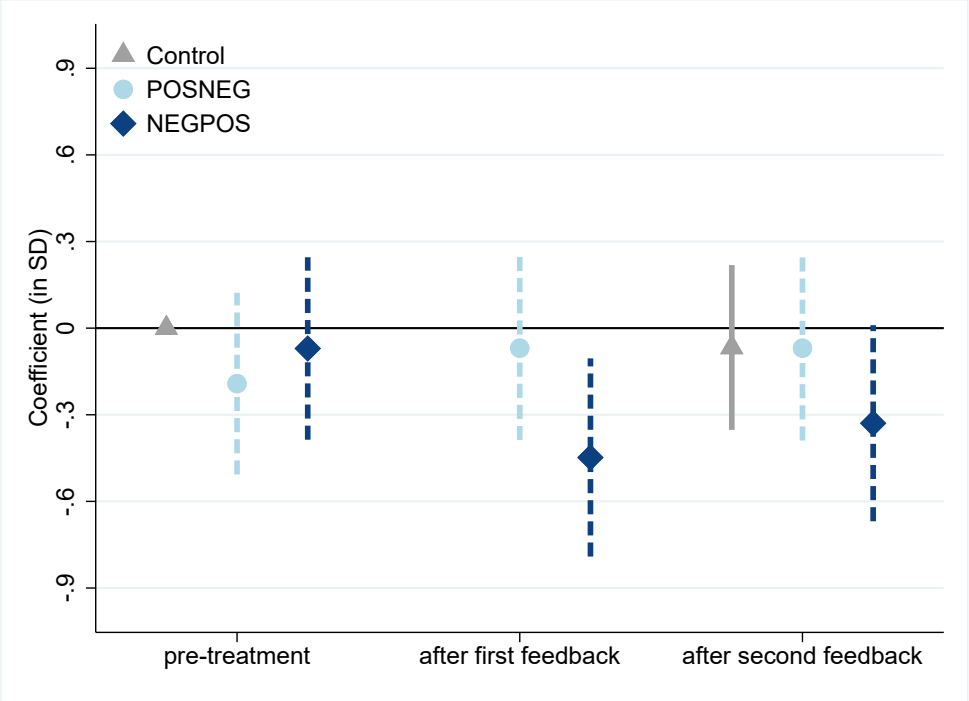
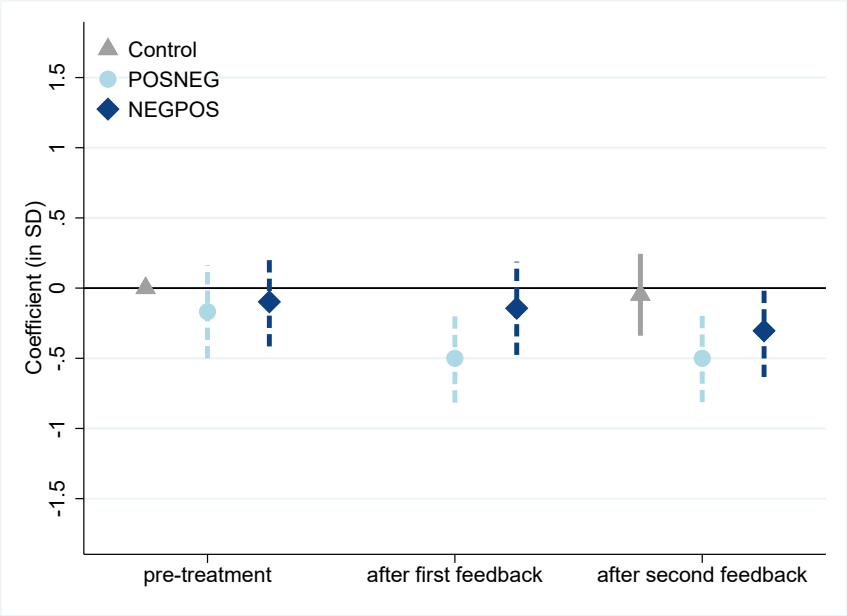


Figure 4.2 : Evolution of Motivation to Study for the Exam



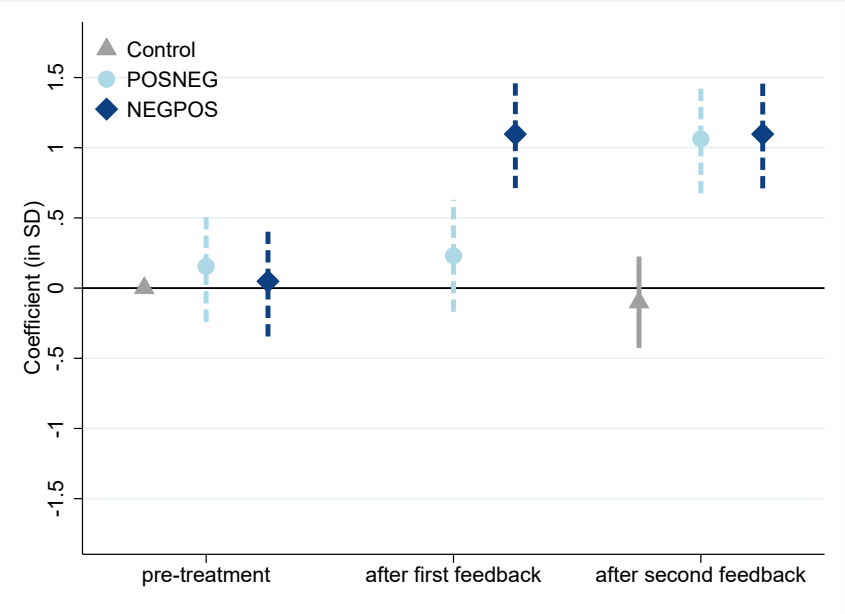
Notes: Plot of coefficients corresponding to equation (4.2) with motivation as an outcome, including the respective 95% confidence intervals. The corresponding regression table can be found in appendix table A4.6.

Figure 4.3 : Evolution of whether Positive Feedback Topic is on Study List



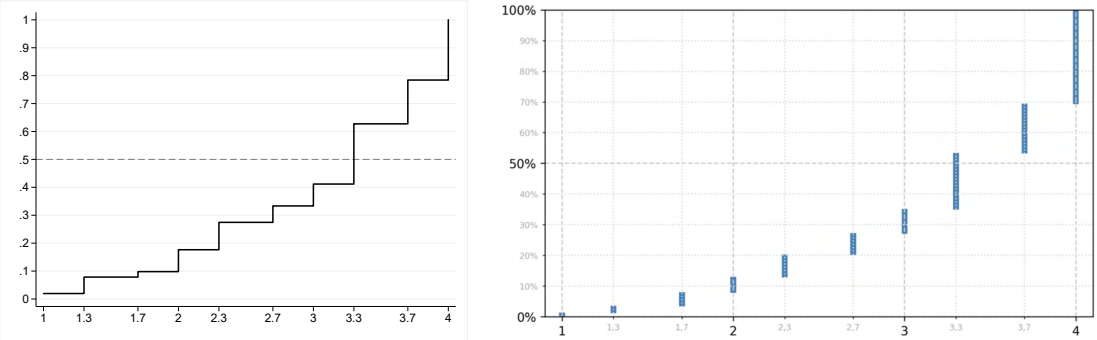
Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the respective 95% confidence intervals. The corresponding regression table can be found in appendix table A4.9.

Figure 4.4 : Evolution of whether Negative Feedback Topic is on Study List



Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the respective 95% confidence intervals. The corresponding regression table can be found in appendix table A4.11.

Figure 4.5 : Selection into Grade Reporting (Course I)

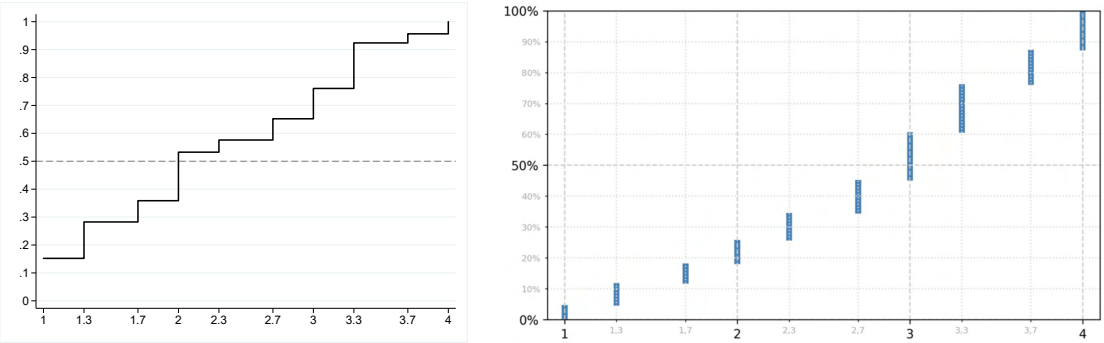


(A) Cumulative distribution function of grades in experiment

(B) Official cumulative distribution function (all students)

Notes: Cumulative distribution functions of exam grades in course I for the experimental sample and all students. Data source panel B: official university statistics.

Figure 4.6 : Selection into Grade Reporting (Course II)

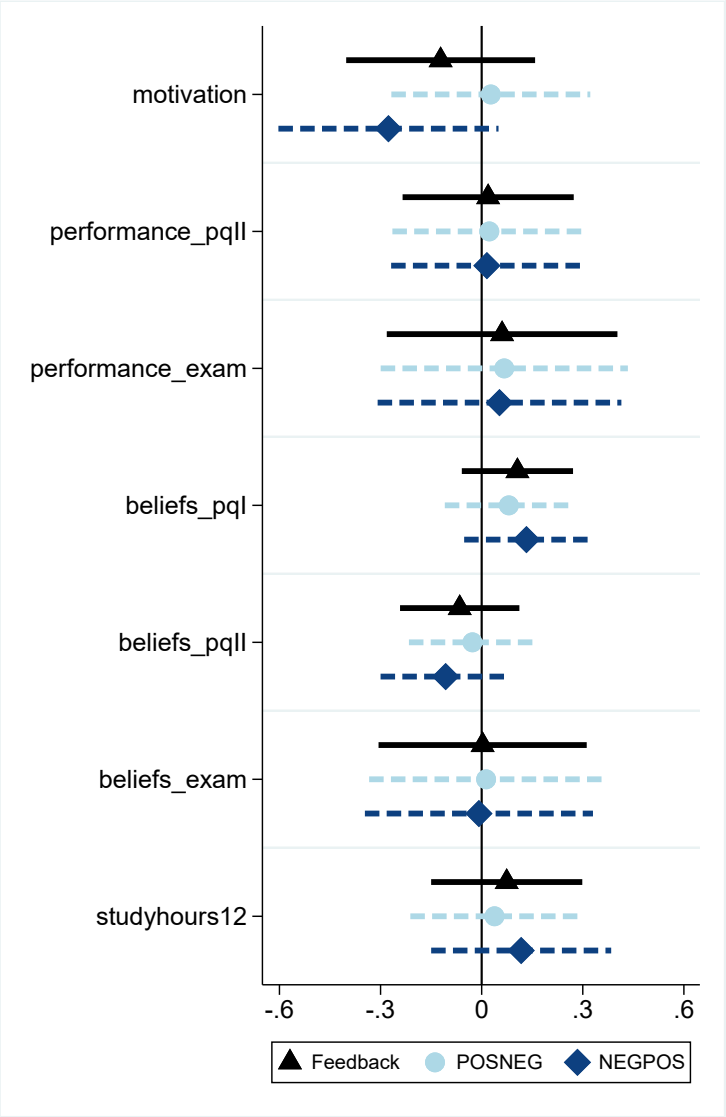


(A) Cumulative distribution function of grades in experiment

(B) Official cumulative distribution function (all students)

Notes: Cumulative distribution functions of exam grades in course II for the experimental sample and all students. Data source panel B: official university statistics.

Figure 4.7 : Effects of Feedback vs no Feedback



Notes: Plot of coefficients corresponding to equation (4.1) with dummies for being treated in general (any of the sequences, *Feedback*) or in one of the respective treatment groups (*POSNEG* and *NEGPOS*) compared to the control group, including the respective 95% confidence intervals. Dependent variables as described in the legend, all outcomes are standardized across all groups. OLS regressions. All regressions contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on, pre-treatment outcome and performance, and all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of all treatment and control groups.

Table 4.1 : Overview of the Feedback Questionnaire

group	share	feedback I	**motivation II** *beliefs II* *study plan II*	feedback II	**motivation III** *beliefs III* *study plan III*	practice quest. II (**performance II**)
T1	0.4	positive	x	negative	x	total points
T2	0.4	negative	x	positive	x	total points
C	0.2	no feedback		no feedback	x	total points

Notes: Group shares and ordering of elements in the feedback questionnaire.

Table 4.2 : Number of Observations by Stage of the Experiment and Course

	Course size (1)	Sign-up (2)	PQI (3)	FB + PQII (4)	Without Duplicates (5)	Post-Exam w/o dupl. (6)	Grades w/o dupl. (7)
Course I							
Observations	644	132	101	86	86	81	66
% of previous stage		20.5	76.52	85.15	100	94.19	76.74
% of initial obs.		100.00	76.52	65.15	65.15	61.36	50
Course II							
Observations	1006	227	165	145	139	129	103
% of previous stage		22.56	72.69	87.88	95.86	88.96	74.1
% of initial obs.		100.00	72.69	63.88	61.23	56.83	45.37
Total							
Observations	1650	359	266	231	225	210	169
% of previous stage		21.76	74.09	86.84	97.40	93.33	75.11
% of initial obs.		100.00	74.09	64.34	62.67	58.49	47.07

Notes: Number of observations and relative shares of participants at all experimental stages, from sign-up to grade reporting, by course. The course size refers to the number of students who took the exam on the first available date. Column (5) excludes the observation for the second course of all students who participated in the experiment for both courses. Data source col. 1: official university statistics.

Table 4.3 : Number of Observations by Stage of the Experiment and Treatment Status

	(1) PQI	(2) FB + PQII	(3) Without Duplicates	(4) Post-Exam w/o dupl.	(5) Grades w/o dupl.
Control Group					
Observations	51	44	44	42	29
% of previous stage		86.27	100	95.45	65.91
% of initial obs.	100	86.27	85.27	82.35	56.86
POSNEG					
Observations	106	91	89	81	69
% of previous stage		85.85	97.80	91.01	77.53
% of initial obs.	100	85.85	83.96	76.41	65.09
NEGPOS					
Observations	109	96	92	87	71
% of previous stage		83.49	95.83	94.56	77.17
% of initial obs.	100	83.49	84.40	79.82	65.14

Notes: Number of observations and relative shares of participants from the first practice questions to grade reporting, by treatment status. Column (3) excludes the observation for the second course of all students who participated in the experiment for both courses.

Table 4.4 : Descriptive Statistics

	Obs.	Mean	Median	Std. Dev.	Min.	Max.
Female	225	0.56	1.00	0.50	0	1
Age	224	20.18	20.00	3.00	18	54
Business Administration	225	0.56	1.00	0.50	0	1
Economics	225	0.17	0.00	0.38	0	1
Bus. and Econ. Education	225	0.06	0.00	0.23	0	1
Minor Bus./Econ./Education	225	0.12	0.00	0.33	0	1
Other major	225	0.10	0.00	0.30	0	1
Semester	225	1.93	1.00	1.21	1	7
German high school degree	225	0.90	1.00	0.30	0	1
High school GPA	202	1.74	1.70	0.46	1	3
Last math grade	201	1.85	2.00	0.87	1	5
Mother university degree	224	0.54	1.00	0.50	0	1
Father university degree	225	0.63	1.00	0.48	0	1
Mother employed	224	0.85	1.00	0.36	0	1
Father employed	225	0.87	1.00	0.34	0	1
First-generation migrant	225	0.16	0.00	0.37	0	1
Second-generation migrant	225	0.20	0.00	0.40	0	1
Patience	225	3.56	3.67	0.69	2	5
Risk aversion	225	3.03	3.00	0.94	1	5
Conscientiousness	225	3.83	4.00	0.78	2	5
Neuroticism	225	3.04	3.00	1.06	1	5
Openness	225	3.46	3.50	1.07	1	5
Extraversion	225	3.39	3.50	1.12	1	5
Agreeableness	225	3.22	3.00	0.88	1	5
Feedback aversion	225	2.10	2.00	0.74	1	5
Self efficacy	225	3.89	4.00	0.69	1	5
Motivation university studies	225	3.86	4.00	0.96	1	5
Weekly hours invested in course	225	6.43	5.00	9.69	1	135

Notes: Descriptive statistics of control variables used in the later analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table 4.5 : Balancing Check for all Control Variables

Variable	(1) Control	(2) POSNEG	(3) NEGPOS	(4) (2) vs (1)	(5) (3) vs (1)	(6) (2) vs (3)
Female	0.614	0.562	0.543	-0.052	-0.070	0.018
Age	20.273	20.056	20.253	-0.217	-0.020	-0.197
Business Administration	0.568	0.573	0.533	0.005	-0.036	0.040
Economics	0.227	0.180	0.130	-0.047	-0.097	0.049
Bus. and Econ. Education	0.068	0.090	0.022	0.022	-0.046	0.068**
Minor Bus./Econ./Education	0.068	0.101	0.163	0.033	0.095	-0.062
Other major	0.068	0.056	0.152	-0.012	0.084	-0.096**
Semester	2.045	2.045	1.772	-0.001	-0.274	0.273
German high school degree	0.909	0.899	0.902	-0.010	-0.007	-0.003
High school GPA	1.740	1.725	1.742	-0.015	0.002	-0.017
Last math grade	2.017	1.746	1.873	-0.271*	-0.144	-0.127
Mother university degree	0.591	0.545	0.500	-0.045	-0.091	0.045
Father university degree	0.614	0.618	0.641	0.004	0.028	-0.023
Mother employed	0.909	0.830	0.848	-0.080	-0.061	-0.018
Father employed	0.841	0.865	0.891	0.024	0.050	-0.026
First-generation migrant	0.136	0.202	0.141	0.066	0.005	0.061
Second-generation migrant	0.182	0.169	0.228	-0.013	0.046	-0.060
Patience	3.621	3.543	3.540	-0.078	-0.081	0.003
Risk aversion	3.261	2.955	2.984	-0.306*	-0.278	-0.029
Conscientiousness	3.875	3.865	3.783	-0.010	-0.092	0.083
Neuroticism	3.375	2.871	3.049	-0.504***	-0.326*	-0.178
Openness	3.523	3.399	3.500	-0.124	-0.023	-0.101
Extraversion	3.489	3.483	3.250	-0.005	-0.239	0.233
Agreeableness	3.352	3.169	3.196	-0.184	-0.157	-0.027
Feedback aversion	2.030	2.094	2.149	0.063	0.118	-0.055
Self efficacy	3.947	3.910	3.841	-0.037	-0.106	0.070
Motivation university studies	3.977	3.798	3.870	-0.180	-0.108	-0.072
Weekly hours invested in course	5.500	7.635	5.710	2.135	0.210	1.925
Observations	44	89	92	133	136	181

Notes: Balancing check between control and treatment groups on all controls used in the analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.6 : Treatment Effects on Motivation Between and After Feedback Elements

	After first feedback			After both feedback elements		
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.343**	0.411***	0.434***	0.220	0.292**	0.349**
	(0.151)	(0.128)	(0.136)	(0.154)	(0.134)	(0.136)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment motivation		✓	✓		✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	181	181	179	181	181	179
R^2	0.144	0.366	0.512	0.104	0.347	0.498

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam between feedback elements (col.s 1-3) and after both feedback elements (col.s 4-6). OLS regressions. Dependent variable: (standardized) motivation to study for the respective exam. Col.s 1 and 4 only include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 add 5 pre-treatment motivation and performance. Finally, col.s 3 and 6 add all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table 4.7 : Treatment Effects on Beliefs

	Practice set I		Exam		Practice set II
	after fb1	after fb2	after fb1	after fb2	after fb2
	(1)	(2)	(3)	(4)	(5)
POSNEG	0.501***	-0.025	0.136*	0.058	0.069
	(0.081)	(0.090)	(0.077)	(0.075)	(0.145)
Female × Course Indicator	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓
Pre-treatment expectations	✓	✓	✓	✓	
Pre-treatment performance	✓	✓	✓	✓	✓
Observations	181	181	181	181	181
R^2	0.770	0.678	0.760	0.755	0.115

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on beliefs. OLS regressions. Dependent variable: (standardized) beliefs about performance in practice set 1 (col.s 1 and 2), exam grades (col.s 3 and 4) and practice set 2 (col. 5). All columns include a female × course indicator for each randomization cell, feedback topic dummies, and pre-treatment beliefs (where applicable) and performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table 4.8 : Treatment Effects on Performance

	Practice set II			Exam		
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	-0.074	-0.095	-0.008	0.164	0.099	0.034
	(0.145)	(0.128)	(0.137)	(0.154)	(0.130)	(0.124)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	181	181	179	140	140	139
R^2	0.168	0.327	0.505	0.318	0.526	0.693

Standard errors in parentheses

$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on performance in the second set of practice questions and the exam. OLS regressions. Dependent variable: (standardized) performance points (practice set 2, col.s 1-3) and standardized and inverted exam grades (col-s 4-6). Columns 1 and 4 show estimates without controls but include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 and 5 add pre-treatment performance. Finally, col.s 3 and 6 add all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table 4.9 : Heterogeneities by Exam-feedback Topic Correspondence for Exam Grades

	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG=1	0.123	0.052	-0.002	0.734***	0.633***	0.396**
	(0.212)	(0.163)	(0.170)	(0.219)	(0.172)	(0.165)
Best fb topic not in exam=1	0.809	0.411*	0.472			
	(0.644)	(0.232)	(0.292)			
POSNEG=1 × Best fb topic not in exam=1	0.033	0.068	0.033			
	(0.292)	(0.245)	(0.245)			
Worst fb topic not in exam=1				1.337***	1.390***	1.026***
				(0.494)	(0.392)	(0.352)
POSNEG=1 × Worst fb topic not in exam=1				-1.131***	-1.065***	-0.711***
				(0.290)	(0.235)	(0.259)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	140	140	139	140	140	139
R^2	0.327	0.529	0.696	0.407	0.611	0.721

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on performance in the exam. OLS regressions. Dependent variable: (standardized) exam grades. Columns 1 and 4 only include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 and 5 add pre-treatment performance. Finally, col.s 3 and 6 add all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table 4.10 : Feelings after Feedback for Treatment Groups

	(1)	(2)	(3)
Variable	POSNEG	NEGPOS	(2) vs (1)
Feelings after FB I	3.562	2.250	-1.312***
Feelings after FB II	2.404	3.957	1.552***
Observations	89	92	181

Notes: Feelings after receiving each of the feedback elements for the two treatment groups.

Table 4.11 : Descriptive Statistics from the Post-exam Questionnaire

	Obs.	Mean	Median	Std. Dev.	Min.	Max.
Difficulty Exam	210	2.15	2.00	0.90	1	5
Difficulty PQ I	210	3.20	3.00	1.05	1	5
Difficulty PQ II	210	3.87	4.00	0.94	1	5
Usefulness PQ I	210	3.17	3.00	1.24	1	5
Usefulness PQ II	210	2.73	3.00	1.08	1	5
Correct recall FB1 Yes/No	210	0.84	1.00	0.36	0	1
Correct recall FB1 Elements	138	0.88	1.00	0.32	0	1
Correct recall FB1 Ordering	122	0.91	1.00	0.29	0	1
Usefulness Feedback PQ I	141	2.39	2.00	1.13	1	5
Feelings Feedback PQI	141	2.86	3.00	0.80	1	5
Correct recall FB2 Yes/No	210	0.89	1.00	0.32	0	1
Correct recall FB2 Points	186	0.95	1.00	0.22	0	1
Usefulness Feedback PQ II	186	2.64	2.50	1.17	1	5

Notes: Descriptive statistics from the post-exam questionnaire.

Table 4.12 : Treatment Effects on Overall Feelings about Feedback

	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.791***	0.802***	0.829***	0.836***	0.875***	0.859***
	(0.157)	(0.157)	(0.156)	(0.161)	(0.160)	(0.190)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	138	138	138	138	138	136
R^2	0.157	0.183	0.240	0.241	0.260	0.436

Standard errors in parentheses

$^{\star}\ p < 0.10, ^{\star\star}\ p < 0.05, ^{\star\star\star}\ p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on post-exam overall feelings about the feedback on the first set of practice questions. OLS regressions. Dependent variable: (standardized) feelings about feedback. Col. 1 shows the regression without any controls, col. 2 only includes a female × course indicator for each randomization cell. Col.s 3-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance, respectively. Finally, col. 6 adds all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Appendix

A4.1 Report of Heterogeneous Treatment Effects as Specified in the Pre-Analysis Plan

In this section, I will present more detailed results on all pre-specified dimensions of heterogeneity that from a theoretical perspective might have been of interest. Due to the relatively small sample size, the interpretability of these results is limited, but for completeness, I will present all respective estimations.

Differences by Gender

The main dimension along which I expected to find heterogeneous effects, was the gender of participants. In fact, prior evidence on feedback suggests that women react more strongly to feedback, especially to negative elements (Buser, 2016; Coffman et al., 2021; Goulas and Megalokonomou, 2021). Interestingly in my case, no differences by gender can be observed for most of the outcomes, including motivation, as can be seen in table A4.23. The only outcome that shows a gender component is exam grades in column 3. Maybe surprisingly, it seems as if only male students benefit from the positive feedback first whereas female students might overall actually be negatively affected. As can be seen in column 7, this does not operate through differential effects on study hours for the last two days before the exam.[1]

Furthermore, figures A4.6 and A4.7 in appendix A4.2 show that this can hardly be explained by differences in reactions to feedback topics. Male students do seem to react slightly more optimistically to the positive feedback topic, i.e. are even more likely to take it away from their study plan than females. This is especially the case after the second (positive) feedback element for the NEGPOS group where women almost don't react at all to the positive feedback topic whereas men do. In fact, when running a regression according to equation (4.2) adding female interaction terms for all variables, all interaction terms with dummies from the NEGPOS group are positive but not significant, indicating that women are less likely to remove the positive feedback topic from their study plan in this treatment group (not shown). This is unfortunately also the case for the dummies indicating the initial share of females and males who had that topic on their study list in the NEGPOS group which poses a caveat on the above interpretation, especially since the coefficient on this interaction term is actually statistically significant. For negative feedback topics, eyeballing the graphs would suggest a similar pattern where women seem to be more reactive to the negative feedback independently of when they received it. Econometrically though, none of the respective dummy×female interaction terms is statistically significant (not shown).

[1] All regressions in this section follow the preferred specification of my main analyses without the full set of controls. Results in general look very similar when including all controls from table 4.4.

Differences by Initial Performance Level

Another dimension of interest is students' initial performance level because of the prior evidence on differences in reactions to feedback between low- and high-achievers (Bandiera et al., 2015; Goulas and Megalokonomou, 2021; Hermes et al., 2021). Indeed, one goal of this paper was to be able to look at differential effects on low- and high-achievers. This is usually not possible when feedback is given relative to others, i.e. where low- (high-) achievers naturally always receive negative (positive) feedback only. In my setting, due to the nature of the feedback as a within-person performance comparison, I can instead look at the effects of early positive feedback on low-achievers. Table A4.24 shows treatment effects on individuals depending on their performance in the first set of practice questions which took place before treatment. The only outcomes for which this comparison seems to matter after both feedback elements were received, are motivation and beliefs about immediate performance. The negative coefficient on the interaction term between receiving positive feedback first and the respective performance points suggests that for better-performing students the (beneficial) effect of positive feedback first fades with increasing performance.

Again, this can be explored more in detail when looking at the evolution of motivation compared to the control group receiving no feedback on the first set of practice questions. Figure A4.8 shows that the pattern observed in figure 4.2 is only clearly visible for those with below-median performance on the first set of practice questions. Instead, for high-achievers there is a drop in motivation for individuals from the POSNEG group as well. This could have two reasons: either this just reflects a general trend of falling motivation for high-achievers when moving closer to the exam or high-achievers are demotivated by positive and negative feedback equally. The fact that we don't see a clear drop for the control group rather speaks in favor of the second explanation, even though there is no significant difference in feedback aversion as measured pre-treatment between these two groups (not shown). From a fully saturated regression, I can also infer that the coefficients on POSNEG_fb1 and POSNEG_fb2 are statistically significantly different for the two performance groups. These results again hint towards a ceiling effect for motivation that might in this case be especially pronounced for high-performers. In fact, initial average levels of motivation to study for the respective exam were significantly higher for those with later above-median performance in the practice questions (not shown).

Differences by Socio-economic Background

A further dimension of heterogeneity that could theoretically be of interest, is the socio-economic background (SES) of the respective students. Students' socio-economic status could matter for their reaction to the ordering of positive and negative feedback elements if the (order of) feedback they are used to receiving is connected to the parents' academic background. For example, one might assume that parents without academic background might be more likely to emphasize positive feedback regarding their children's performance

in the academic context given that it exceeds their own highest level of education and hence might also be more likely to mention this feedback element first.

In this paper, I measure SES as having at least one parent with a university degree (high SES) compared to none (low SES). Table A4.25 shows a regression including an interaction term for the treatment variable and an individual's SES as measured here. The general picture suggests that for individuals with no academic parent, receiving positive feedback first is less relevant than for those with academic parents. This is especially credibly visible for beliefs where the coefficient on the interaction term is statistically significant for the first set of practice questions and the exam.

These findings point towards some structurally different ways of how students with (non-) academic parents react to feedback and in particular the ordering of positive and negative elements. One explanation for this could be that these students are particularly highly selected, i.e. only the very well-performing students from this 'low-SES' background can be found in my sample. Alternatively, there could be a difference in how feedback dynamics operate in their families as described above. With my data, I cannot make any statement about the second channel, but I can indeed have a closer look at these students' background characteristics.

In fact, students with different socio-economic backgrounds seem to be systematically different on some characteristics as can be seen from table A4.19. Students with 'low SES' as measured in this setting tend to be older and in higher semesters, they choose different majors and have worse high-school outcomes (on a German scale). Most interestingly, they markedly differ with respect to their personality traits, i.e. individuals with non-academic parents are less conscientious, i.e. more lazy/less thorough, and more neurotic, i.e. more stressed and anxious, or at least perceive themselves in this way when answering the questionnaire. Finally, they tend to be less motivated to study for university in general, but have invested more hours thus far in studying for this course. These pieces of evidence suggest that rather the second mechanism described above is what can explain differences in these students' reactions to feedback ordering since the 'low-SES' students from this study are not particularly highly selected.

Differences by Personality Traits

Similar analyses can be run with patience, risk-aversion, the big 5 personality traits, feedback aversion, and self-efficacy.Figure A4.9 shows a plot for the coefficients on the interaction term of the treatment dummy and the respective trait from regressions of all presented outcomes on all mentioned traits.

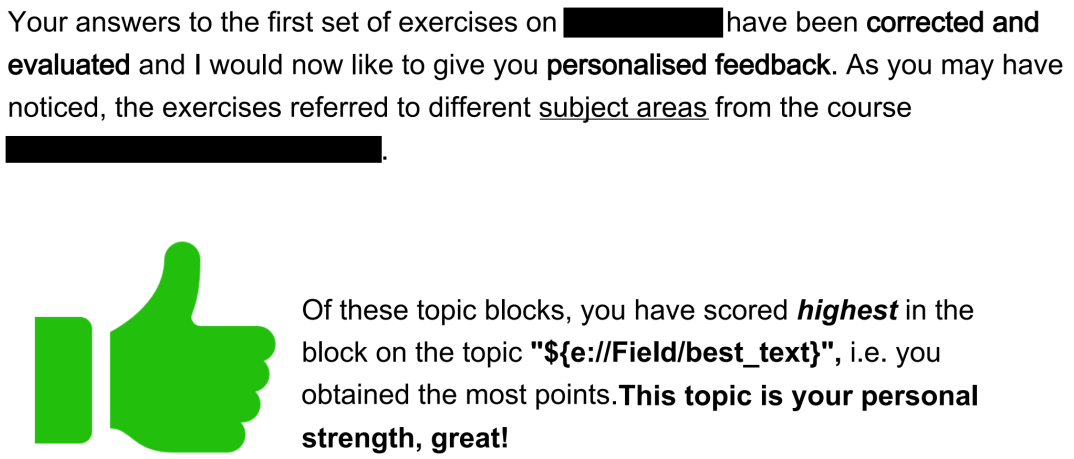
Overall, there is no clear pattern of any trait influencing treatment effects into a certain direction, but there are some exceptions that might be worth looking into. For example, risk aversion seems to play a role for the treatment effects on exam grades: the more risk-averse students are, the less important it is for them to receive the positive feedback element

first (while more risk-averse students in general have higher grades, not shown). Similarly, more neurotic students benefit less from first receiving positive feedback first for their exam performance. On the other hand, neurotic students benefit more from the *positive-negative* feedback order for beliefs about their immediate performance in the second set of practice questions which suggests that neuroticism operates very differently for intuitive reactions compared to medium-term processes. This seems to hold true for openness as well: more open students benefit more from the *positive-negative* ordering both for motivation as well as for their immediate performance, but they also increase their study hours on the last two days before the exam less compared to less open students.

Lastly, self-efficacy plays a role for the treatment effect of the feedback sequence on beliefs about immediate performance after the feedback: more self-efficacious students benefit less from first receiving positive feedback in having higher post-feedback beliefs about their performance in the second set of practice questions (overall, self-efficacy is positively related to immediate performance beliefs, not shown). Maybe surprisingly, no clear patterns can be observed for patience, conscientiousness, extraversion, agreeableness, and feedback aversion.

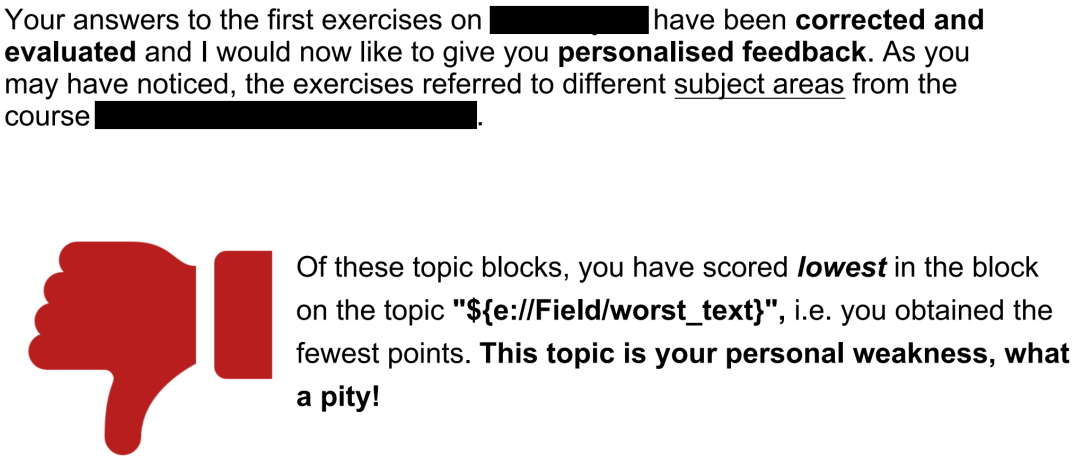
A4.2 Appendix Tables and Figures

Figure A4.1 : Example of Positive Feedback (Translated from German)



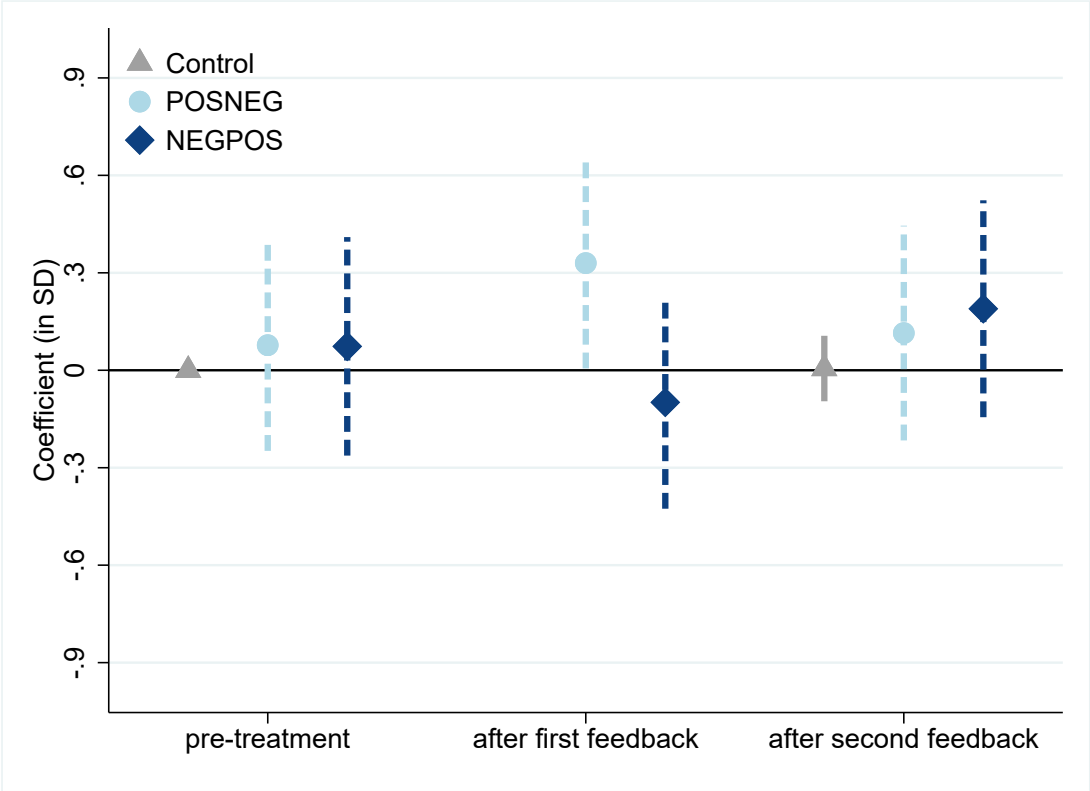
Notes: Screenshot from the feedback questionnaire. Details on the specific course and date have been blackened.

Figure A4.2 : Example of Negative Feedback (Translated from German)



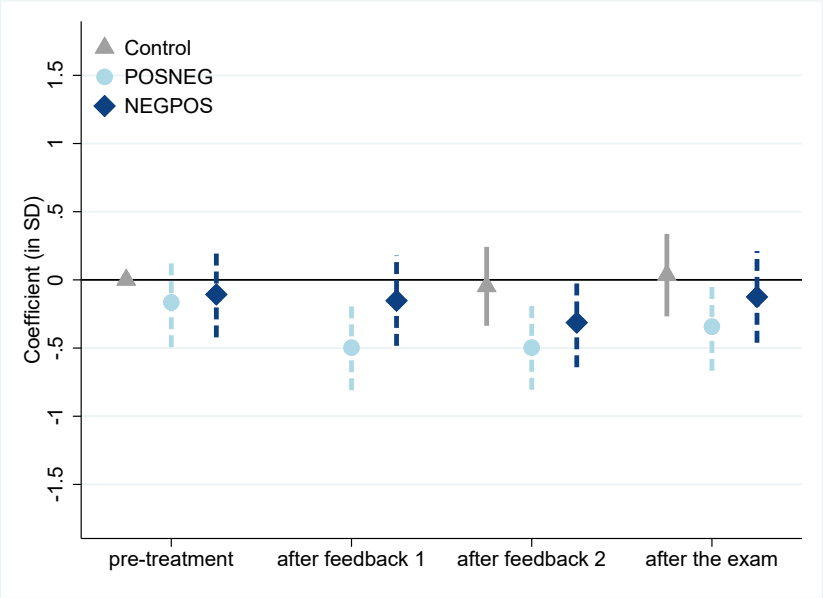
Notes: Screenshot from the feedback questionnaire. Details on the specific course and date have been blackened.

Figure A4.3 : Evolution of Performance Beliefs about PQI



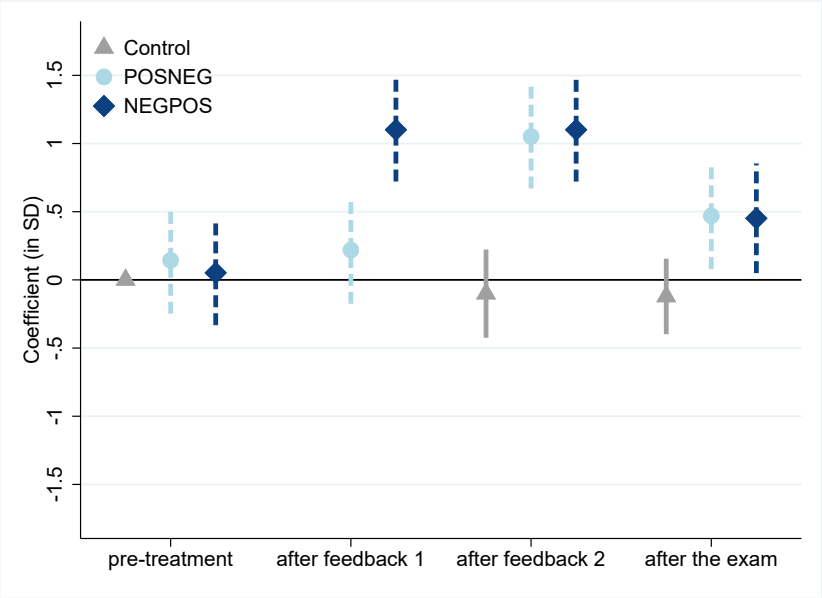
Notes: Plot of coefficients corresponding to equation (4.2) with standardized beliefs about performance in the first set of practice questions, including the respective 95% confidence intervals. The corresponding regression table can be found in appendix table A4.7.

Figure A4.4 : Evolution of whether Positive Feedback Topic is on Study List, incl. Post-exam



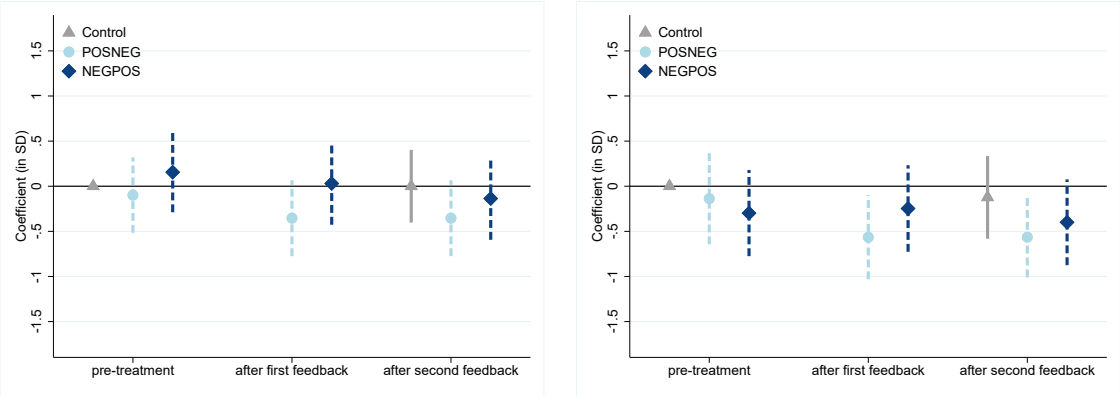
Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the respective 95% confidence intervals. This version is enriched by the students' assessment after the exam. The corresponding regression table can be found in appendix table A4.10.

Figure A4.5 : Evolution of whether Negative Feedback Topic is on Study List, incl. Post-exam



Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the respective 95% confidence intervals. This version is enriched by the students' assessment after the exam. The corresponding regression table can be found in appendix table A4.12.

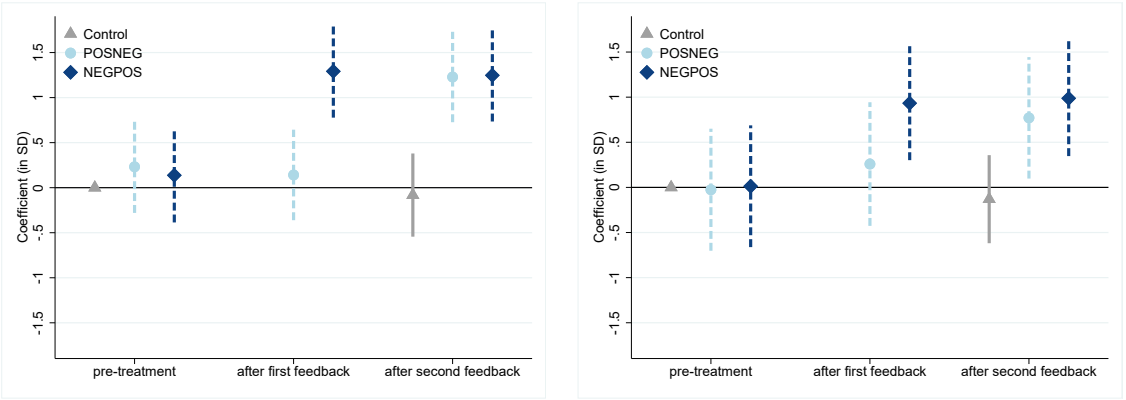
Figure A4.6 : Evolution of whether Positive Feedback Topic is on Study List, by Gender



(A) Females (B) Males

Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the respective 95% confidence intervals, by gender. The corresponding regression table can be found in appendix tables A4.13 and A4.14.

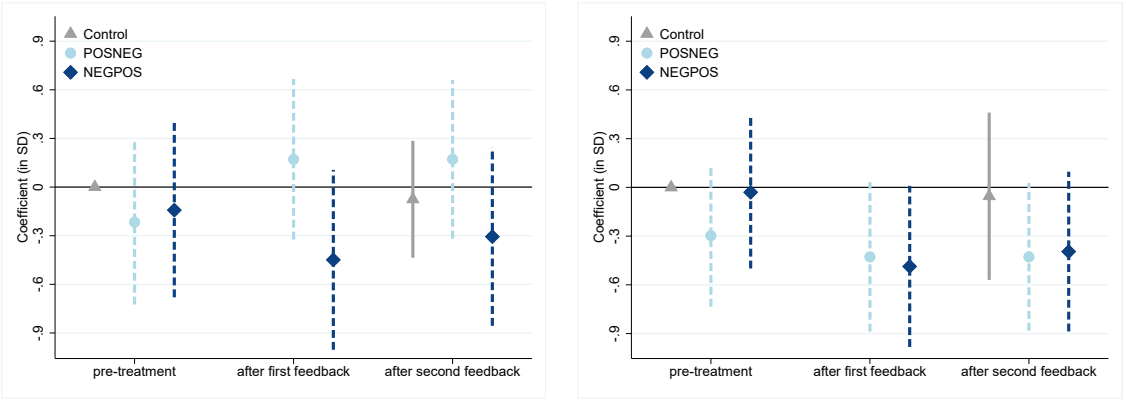
Figure A4.7 : Evolution of whether Negative Feedback Topic is on Study List, by Gender



(A) Females (B) Males

Notes: Plot of coefficients corresponding to equation (4.2) with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the respective 95% confidence intervals, by gender. The corresponding regression table can be found in appendix tables A4.15 and A4.16.

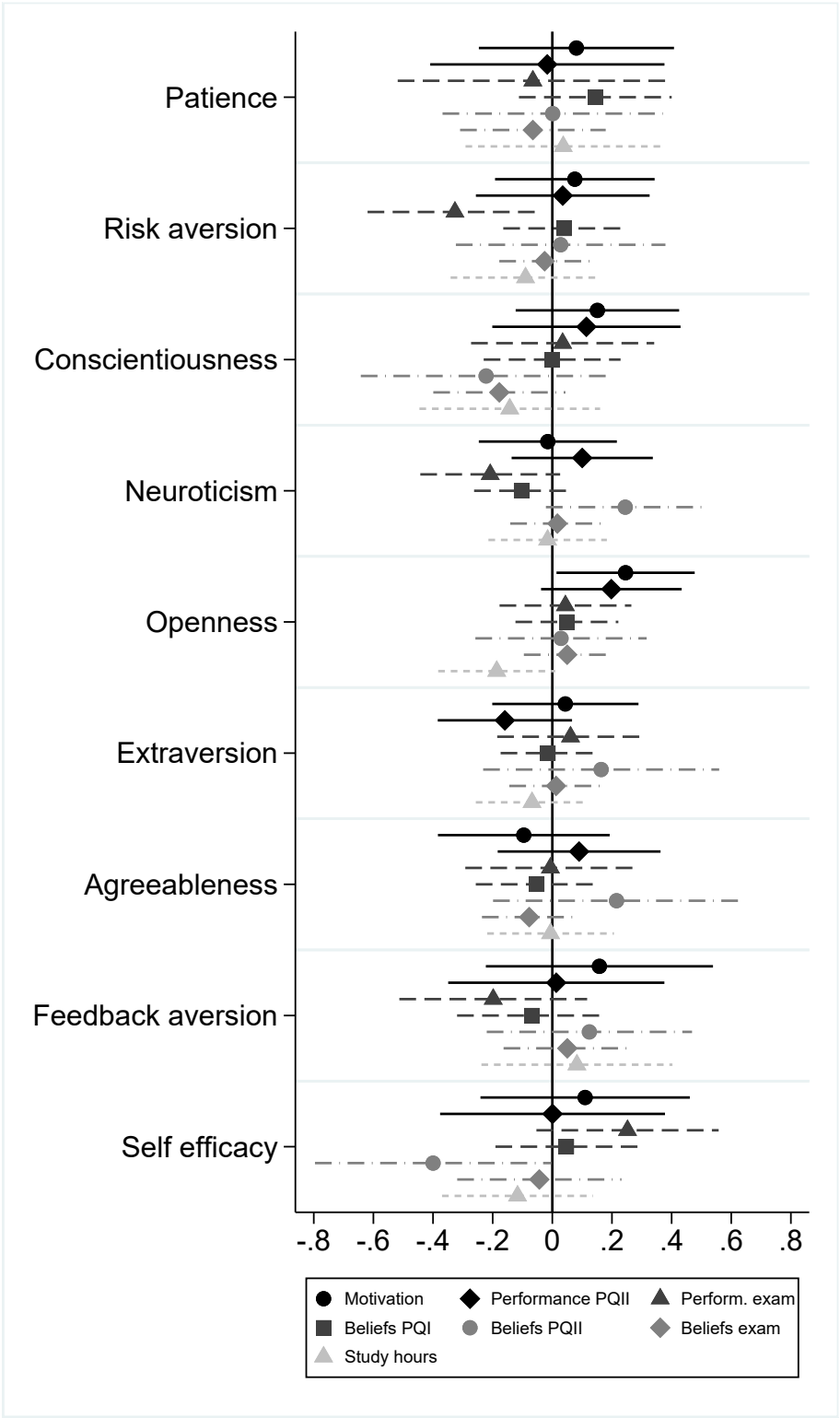
Figure A4.8 : Evolution of Motivation for Pre-treatment Performance Below-/above Median



(A) Below-median pre-treatment performance (B) Above-median pre-treatment performance

Notes: Plot of coefficients corresponding to equation (4.2) with motivation as an outcome, including the respective 95% confidence intervals, by pre-treatment performance above or below the sample median. The corresponding regression tables can be found in appendix tables A4.17 and A4.18.

Figure A4.9 : Plot of all Coefficients on Personality Trait Interactions for all Outcomes



Notes: Coefficient estimates on the interaction term between a treatment dummy and the respective personality trait, including the respective 95% confidence intervals. Dependent variables as described in the legend, all outcomes are standardized within the treatment groups. OLS regressions. All regressions contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and pre-treatment outcomes and performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.1 : Descriptive Statistics for Pre-treatment Outcomes

	Obs.	Mean	Median	Std. Dev.	Min.	Max.
Motivation study exam	225	3.54	4.00	1.06	1	5
Points PQ I	225	8.46	8.50	3.82	0	18
Before answering						
Expected points PQ I	225	12.60	13.00	3.58	1	20
Self evaluation exp. pts PQI	225	2.87	3.00	1.05	1	5
Expected prob. Q1 (PQI)	225	17.72	10.00	21.47	0	100
Expected prob. Q2 (PQI)	225	30.46	30.00	20.01	0	100
Expected prob. Q3 (PQI)	225	32.88	30.00	20.04	0	90
Expected prob. Q4 (PQI)	225	18.94	10.00	23.21	0	100
After answering						
Expected points PQ I	225	8.72	8.00	4.27	0	18
Self evaluation exp. pts PQI	225	1.89	2.00	0.92	1	5
Expected prob. Q1 (PQI)	225	37.64	30.00	33.48	0	100
Expected prob. Q2 (PQI)	225	33.57	35.00	23.11	0	100
Expected prob. Q3 (PQI)	225	20.94	15.00	21.10	0	80
Expected prob. Q4 (PQI)	225	7.84	0.00	16.86	0	100
Expected grade (German scale)	225	2.31	2.30	0.69	1	4
Self evaluation exp. grade	225	3.22	3.00	1.06	1	5
Expected prob. Q1 (grade)	225	12.92	5.00	18.56	0	100
Expected prob. Q2 (grade)	225	25.35	25.00	18.43	0	80
Expected prob. Q3 (grade)	225	36.27	35.00	20.31	0	100
Expected prob. Q4 (grade)	225	25.46	10.00	28.42	0	100
Planned study hours (-6)	225	2.44	2.00	1.92	0	12
Planned study hours (-5)	225	2.48	2.00	1.89	0	10
Planned study hours (-4)	225	2.13	2.00	1.83	0	8
Planned study hours (-3)	225	2.04	2.00	1.80	0	12
Planned study hours (-2)	225	3.14	3.00	2.45	0	12
Planned study hours (-1)	225	4.21	4.00	2.47	0	12
Positive fb topic on study list	225	0.27	0.00	0.44	0	1
Negative fb topic on study list	225	0.29	0.00	0.46	0	1

Notes: Descriptive statistics of all pre-treatment outcomes. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table A4.2 : Balancing Check for Pre-treatment Outcomes

Variable	(1) Control	(2) POSNEG	(3) NEGPOS	(4) (2) vs (1)	(5) (3) vs (1)	(6) (2) vs (3)
Motivation study exam	3.591	3.461	3.598	-0.130	0.007	-0.137
Points PQ I	7.852	8.584	8.636	0.732	0.784	-0.052
Before answering						
Expected points PQ I (pre)	12.136	12.910	12.511	0.774	0.375	0.399
Self evaluation exp. pts PQI (pre)	2.773	2.955	2.826	0.182	0.053	0.129
Expected prob. Q1 (PQI) (pre)	24.182	14.079	18.152	-10.103***	-6.030	-4.074
Expected prob. Q2 (PQI) (pre)	29.295	29.157	32.272	-0.138	2.976	-3.114
Expected prob. Q3 (PQI) (pre)	31.045	37.303	29.489	6.258*	-1.556	7.814***
Expected prob. Q4 (PQI) (pre)	15.477	19.461	20.087	3.983	4.610	-0.626
After answering						
Expected points PQ I (post)	7.932	9.090	8.728	1.158	0.796	0.362
Self evaluation exp. pts PQI (post)	1.886	1.933	1.859	0.046	-0.028	0.074
Expected prob. Q1 (PQI) (post)	40.227	33.764	40.163	-6.463	-0.064	-6.399
Expected prob. Q2 (PQI) (post)	33.523	37.056	30.228	3.533	-3.294	6.828**
Expected prob. Q3 (PQI) (post)	18.955	23.506	19.402	4.551	0.448	4.103
Expected prob. Q4 (PQI) (post)	7.295	5.674	10.207	-1.621	2.911	-4.532*
Expected grade (German scale)	2.486	2.275	2.254	-0.211*	-0.232*	0.021
Self evaluation exp. grade	3.159	3.157	3.304	-0.002	0.145	-0.147
Expected prob. Q1 (grade)	17.364	10.315	13.315	-7.049**	-4.048	-3.001
Expected prob. Q2 (grade)	26.750	24.910	25.109	-1.840	-1.641	-0.199
Expected prob. Q3 (grade)	35.227	40.067	33.098	4.840	-2.129	6.970**
Expected prob. Q4 (grade)	20.659	24.708	28.478	4.049	7.819	-3.770
Planned study hours (-6)	2.455	2.669	2.223	0.214	-0.232	0.446
Planned study hours (-5)	2.557	2.562	2.353	0.005	-0.204	0.209
Planned study hours (-4)	1.739	2.185	2.261	0.447	0.522	-0.075
Planned study hours (-3)	1.841	2.247	1.937	0.406	0.096	0.310
Planned study hours (-2)	2.591	3.163	3.370	0.572	0.779*	-0.207
Planned study hours (-1)	3.932	4.208	4.337	0.276	0.405	-0.129
Positive fb topic on study list	0.341	0.213	0.283	-0.127	-0.058	-0.069
Negative fb topic on study list	0.273	0.326	0.272	0.053	-0.001	0.054
Observations	44	89	92	133	136	181

Notes: Balancing check between control and treatment groups on all pre-treatment outcomes. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table A4.3 : Treatment Effects on Motivation with LASSO Regressions

	LASSO		Double LASSO	
	(restricted)	(unrestricted)	(restricted)	(unrestricted)
	(1)	(2)	(3)	(4)
POSNEG	0.290*	0.236	0.298*	0.244
	(0.129)	(0.121)	(0.122)	(0.122)
Observations	179	179	179	179

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam. LASSO (*lasso* in STATA, col.s 1-2) and double LASSO (*dsregress* in STATA, col.s 3-4) regressions with selection of covariates and default parameters. Dependent variable: (standardized) motivation to study for the respective exam. Columns 1 and 3 select choice of controls on those described in table 4.4 whereas col.s 2 and 4 also allow to select among the female × course indicator, feedback topic dummies, and pre-treatment motivation and performance. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.4 : Treatment Effects on Motivation between Feedback Elements

	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.302**	0.305**	0.343**	0.399***	0.411***	0.434***
	(0.147)	(0.147)	(0.151)	(0.130)	(0.128)	(0.136)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.023	0.045	0.144	0.343	0.366	0.512

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam between feedback elements. OLS regressions. Dependent variable: (standardized) motivation to study for the exam. Col. 1 shows the regression without any controls, col. 1 only includes a female × course indicator for each randomization cell. Col.s 2-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance, respectively. Finally, col. 6 adds all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.5 : Treatment Effects on Motivation after both Feedback Elements

	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.201	0.202	0.220	0.278**	0.292**	0.349**
	(0.148)	(0.148)	(0.154)	(0.135)	(0.134)	(0.136)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.010	0.029	0.104	0.315	0.347	0.498

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam after both feedback elements. OLS regressions. Dependent variable: (standardized) motivation to study for the exam. Col. 1 shows the regression without any controls, col. 1 only includes a female × course indicator for each randomization cell. Col.s 2-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance, respectively. Finally, col. 6 adds all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.6 : Evolution of Motivation to Study for the Exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.192	-0.180	-0.273*
	(0.160)	(0.159)	(0.164)
After fb1 POSNEG	-0.069	-0.057	-0.146
	(0.162)	(0.160)	(0.168)
After fb2 POSNEG	-0.069	-0.057	-0.171
	(0.162)	(0.161)	(0.168)
Initial NEGPOS	-0.070	-0.058	-0.066
	(0.160)	(0.160)	(0.166)
After fb1 NEGPOS	-0.448**	-0.436**	-0.437**
	(0.174)	(0.174)	(0.180)
After fb2 NEGPOS	-0.329*	-0.317*	-0.317*
	(0.172)	(0.172)	(0.177)
After-treatment control	-0.067	-0.067	-0.074
	(0.145)	(0.144)	(0.138)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.408	0.405	0.406

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 4.2 using equation (4.2). Dependent variable: (standardized) motivation to study for the respective exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.7 : Evolution of Performance Beliefs about PQI

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.077	0.112	0.004
	(0.165)	(0.166)	(0.173)
After fb1 POSNEG	0.330**	0.364**	0.268
	(0.165)	(0.164)	(0.176)
After fb2 POSNEG	0.115	0.149	0.046
	(0.168)	(0.167)	(0.179)
Initial NEGPOS	0.074	0.089	-0.013
	(0.171)	(0.174)	(0.180)
After fb1 NEGPOS	-0.099	-0.083	-0.173
	(0.166)	(0.170)	(0.175)
After fb2 NEGPOS	0.189	0.205	0.081
	(0.169)	(0.172)	(0.178)
After-treatment control	0.005	0.005	-0.006
	(0.051)	(0.051)	(0.056)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.430	0.408	0.446

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure A4.3 using equation (4.2). Dependent variable: (standardized) performance beliefs about practice questions 1. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Col. 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, col. 2 excludes high-school performance as a control, and col. 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.8 : Treatment Effects on Study Hours Days 1-2 before the Exam

	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	-0.122	-0.120	-0.113	-0.045	-0.043	-0.073
	(0.148)	(0.143)	(0.144)	(0.102)	(0.103)	(0.120)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment study plan				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.004	0.089	0.124	0.572	0.574	0.661

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on total study hours one and two days before the exam. OLS regressions. Dependent variable: (standardized) total hours students planned to study for the respective exam on the last two days before the exam. Col. 1 shows the regression without any controls, col. 2 only includes a female × course indicator for each randomization cell. Col.s 3-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment planned study hours and performance, respectively. Finally, col. 6 adds all respective controls as can be seen in table 4.4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.9 : Evolution of Positive Feedback Topics

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.168	-0.165	-0.165
	(0.169)	(0.167)	(0.186)
After fb1 POSNEG	-0.500***	-0.496***	-0.544***
	(0.160)	(0.159)	(0.178)
After fb2 POSNEG	-0.500***	-0.496***	-0.544***
	(0.159)	(0.157)	(0.176)
Initial NEGPOS	-0.098	-0.098	-0.198
	(0.161)	(0.161)	(0.175)
After fb1 NEGPOS	-0.144	-0.144	-0.224
	(0.169)	(0.168)	(0.183)
After fb2 NEGPOS	-0.304*	-0.304*	-0.376**
	(0.167)	(0.166)	(0.182)
After-treatment control	-0.047	-0.047	-0.104
	(0.148)	(0.148)	(0.153)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.329	0.328	0.343

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 4.3 using equation (4.2). Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.10 : Evolution of Positive Feedback Topics, incl. Post-exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.165	-0.166	-0.163
	(0.167)	(0.165)	(0.184)
After fb1 POSNEG	-0.497***	-0.497***	-0.542***
	(0.159)	(0.157)	(0.175)
After fb2 POSNEG	-0.497***	-0.497***	-0.542***
	(0.157)	(0.155)	(0.174)
Post-exam POSNEG	-0.343**	-0.344**	-0.357*
	(0.165)	(0.163)	(0.183)
Initial NEGPOS	-0.107	-0.110	-0.207
	(0.160)	(0.159)	(0.173)
After fb1 NEGPOS	-0.153	-0.156	-0.233
	(0.168)	(0.167)	(0.182)
After fb2 NEGPOS	-0.313*	-0.316*	-0.385**
	(0.166)	(0.166)	(0.181)
Post-exam NEGPOS	-0.125	-0.127	-0.186
	(0.171)	(0.170)	(0.185)
After-treatment control	-0.047	-0.047	-0.104
	(0.147)	(0.147)	(0.152)
Post-exam control	0.035	0.035	0.027
	(0.153)	(0.153)	(0.167)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	833	833	743
R^2	0.329	0.328	0.345

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure A4.4 using the extended version of equation (4.2). Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.11 : Evolution of Negative Feedback Topics

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.155	0.139	0.202
	(0.201)	(0.200)	(0.201)
After fb1 POSNEG	0.230	0.215	0.318
	(0.202)	(0.202)	(0.204)
After fb2 POSNEG	1.063***	1.047***	1.154***
	(0.197)	(0.197)	(0.195)
Initial NEGPOS	0.048	0.038	0.081
	(0.199)	(0.200)	(0.198)
After fb1 NEGPOS	1.097***	1.087***	1.137***
	(0.195)	(0.195)	(0.193)
After fb2 NEGPOS	1.097***	1.087***	1.137***
	(0.196)	(0.196)	(0.194)
After-treatment control	-0.101	-0.101	0.000
	(0.166)	(0.165)	(0.164)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.335	0.333	0.339

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 4.4 using equation (4.2). Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.12 : Evolution of Negative Feedback Topics, incl. Post-exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.144	0.131	0.200
	(0.199)	(0.200)	(0.201)
After fb1 POSNEG	0.219	0.207	0.316
	(0.200)	(0.202)	(0.204)
After fb2 POSNEG	1.052***	1.039***	1.152***
	(0.193)	(0.195)	(0.194)
Post-exam POSNEG	0.470**	0.457**	0.510**
	(0.198)	(0.199)	(0.199)
Initial NEGPOS	0.052	0.047	0.105
	(0.195)	(0.196)	(0.193)
After fb1 NEGPOS	1.101***	1.096***	1.161***
	(0.192)	(0.193)	(0.191)
After fb2 NEGPOS	1.101***	1.096***	1.161***
	(0.193)	(0.193)	(0.191)
Post-exam NEGPOS	0.452**	0.446**	0.484**
	(0.204)	(0.205)	(0.204)
After-treatment control	-0.101	-0.101	-0.000
	(0.164)	(0.164)	(0.163)
Post-exam control	-0.122	-0.123	-0.091
	(0.140)	(0.140)	(0.143)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	833	833	743
R^2	0.301	0.300	0.306

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure A4.5 using the extended version of equation (4.2). Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.13 : Evolution of Positive Feedback Topics, Females only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.099	-0.110	-0.124
	(0.211)	(0.212)	(0.233)
After fb1 POSNEG	-0.354*	-0.365*	-0.437*
	(0.213)	(0.211)	(0.235)
After fb2 POSNEG	-0.354*	-0.365*	-0.437*
	(0.213)	(0.211)	(0.235)
Initial NEGPOS	0.156	0.147	-0.004
	(0.224)	(0.223)	(0.242)
After fb1 NEGPOS	0.031	0.022	-0.093
	(0.231)	(0.230)	(0.254)
After fb2 NEGPOS	-0.136	-0.145	-0.270
	(0.231)	(0.230)	(0.251)
After-treatment control	0.000	0.000	-0.087
	(0.203)	(0.202)	(0.210)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	351	351	309
R^2	0.436	0.433	0.446

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure A4.6 using equation (4.2) for females only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.14 : Evolution of Positive Feedback Topics, Males only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.137	-0.097	-0.171
	(0.255)	(0.250)	(0.274)
After fb1 POSNEG	-0.565**	-0.524**	-0.622**
	(0.235)	(0.226)	(0.253)
After fb2 POSNEG	-0.565**	-0.524**	-0.622**
	(0.224)	(0.218)	(0.241)
Initial NEGPOS	-0.298	-0.276	-0.348
	(0.240)	(0.237)	(0.255)
After fb1 NEGPOS	-0.247	-0.225	-0.288
	(0.242)	(0.244)	(0.254)
After fb2 NEGPOS	-0.399*	-0.378	-0.407
	(0.239)	(0.242)	(0.252)
After-treatment control	-0.123	-0.123	-0.130
	(0.231)	(0.229)	(0.247)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	274	274	248
R^2	0.378	0.367	0.385

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure A4.6 using equation (4.2) for males only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.15 : Evolution of Negative Feedback Topics, Females only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.232	0.199	0.465*
	(0.259)	(0.253)	(0.253)
After fb1 POSNEG	0.141	0.109	0.410
	(0.254)	(0.249)	(0.250)
After fb2 POSNEG	1.229***	1.196***	1.520***
	(0.253)	(0.249)	(0.246)
Initial NEGPOS	0.138	0.116	0.368
	(0.264)	(0.262)	(0.254)
After fb1 NEGPOS	1.292***	1.271***	1.501***
	(0.260)	(0.256)	(0.254)
After fb2 NEGPOS	1.247***	1.226***	1.454***
	(0.259)	(0.255)	(0.253)
After-treatment control	-0.082	-0.082	0.092
	(0.233)	(0.232)	(0.223)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	351	351	309
R^2	0.408	0.403	0.417

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure A4.7 using equation (4.2) for females only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.16 : Evolution of Negative Feedback Topics, Males only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.026	-0.018	-0.131
	(0.341)	(0.337)	(0.344)
After fb1 POSNEG	0.259	0.266	0.169
	(0.345)	(0.342)	(0.341)
After fb2 POSNEG	0.771**	0.779**	0.709**
	(0.339)	(0.336)	(0.337)
Initial NEGPOS	0.013	0.025	-0.042
	(0.340)	(0.328)	(0.359)
After fb1 NEGPOS	0.933***	0.945***	0.909***
	(0.318)	(0.310)	(0.332)
After fb2 NEGPOS	0.987***	1.000***	0.972***
	(0.323)	(0.313)	(0.340)
After-treatment control	-0.131	-0.131	-0.139
	(0.246)	(0.244)	(0.263)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	274	274	248
R^2	0.361	0.359	0.376

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure A4.7 using equation (4.2) for males only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.17 : Evolution of Motivation, Below-median Performers only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.217	-0.233	-0.286
	(0.256)	(0.256)	(0.281)
After fb1 POSNEG	0.171	0.155	0.112
	(0.250)	(0.248)	(0.264)
After fb2 POSNEG	0.171	0.155	0.059
	(0.247)	(0.247)	(0.263)
Initial NEGPOS	-0.142	-0.162	-0.259
	(0.271)	(0.267)	(0.280)
After fb1 NEGPOS	-0.449	-0.468*	-0.543*
	(0.280)	(0.279)	(0.286)
After fb2 NEGPOS	-0.306	-0.325	-0.390
	(0.278)	(0.277)	(0.284)
After-treatment control	-0.075	-0.075	-0.089
	(0.182)	(0.181)	(0.153)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	325	325	290
R^2	0.448	0.445	0.407

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure A4.8 using equation (4.2) only for those with below-median performance in the first set of practice questions. Dependent variable: (standardized) motivation to study for the respective exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.18 : Evolution of Motivation, Above-median Performers only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.297	-0.176	-0.384*
	(0.220)	(0.220)	(0.226)
After fb1 POSNEG	-0.428*	-0.307	-0.506**
	(0.232)	(0.231)	(0.232)
After fb2 POSNEG	-0.428*	-0.307	-0.506**
	(0.229)	(0.230)	(0.229)
Initial NEGPOS	-0.030	0.085	0.164
	(0.237)	(0.231)	(0.233)
After fb1 NEGPOS	-0.487*	-0.371	-0.313
	(0.250)	(0.246)	(0.241)
After fb2 NEGPOS	-0.395	-0.280	-0.234
	(0.248)	(0.245)	(0.239)
After-treatment control	-0.055	-0.055	-0.055
	(0.260)	(0.258)	(0.262)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	300	300	267
R^2	0.539	0.520	0.607

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure A4.8 using equation (4.2) only for those with above-median performance in the first set of practice questions. Dependent variable: (standardized) motivation to study for the respective exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all respective controls as can be seen in table 4.4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the respective missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table A4.19 : Balancing Check for all Control Variables, by SES

Variable	(1) No academic parent	(2) At least one academic parent	(3) (1) vs (2)
Female	0.662	0.525	0.137*
Age	21.369	19.692	1.677***
Business Administration	0.554	0.556	-0.002
Economics	0.092	0.200	-0.108*
Bus. and Econ. Education	0.077	0.050	0.027
Minor Bus./Econ./Education	0.092	0.131	-0.039
Other major	0.185	0.063	0.122***
Semester	2.200	1.825	0.375**
German high school degree	0.908	0.900	0.008
High school GPA	1.825	1.698	0.128*
Last math grade	2.034	1.779	0.255*
Mother university degree	0.000	0.750	-0.750***
Father university degree	0.000	0.881	-0.881***
Mother employed	0.859	0.850	0.009
Father employed	0.877	0.869	0.008
First-generation migrant	0.108	0.188	-0.080
Second-generation migrant	0.246	0.175	0.071
Patience	3.456	3.598	-0.142
Risk aversion	3.038	3.022	0.017
Conscientiousness	3.669	3.900	-0.231**
Neuroticism	3.331	2.925	0.406***
Openness	3.523	3.441	0.082
Extraversion	3.385	3.391	-0.006
Agreeableness	3.262	3.197	0.065
Feedback aversion	2.174	2.075	0.099
Self efficacy	3.944	3.867	0.077
Motivation university studies	3.677	3.938	-0.261*
Weekly hours invested in course	8.877	5.436	3.441**
Observations	65	160	225

Notes: Balancing check between high- and low-SES individuals on all controls used in the analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table A4.20 : Balancing Check for all Control Variables for Course 1, by Grade Reporting

Variable	(1) Grade reported	(2) No grade reported	(3) (1) vs (2)
Female	0.682	0.650	-0.032
Age	21.121	20.350	-0.771
Business Administration	0.576	0.650	0.074
Economics	0.000	0.000	0.000
Bus. and Econ. Education	0.076	0.050	-0.026
Minor Bus./Econ./Education	0.152	0.200	0.048
Other major	0.197	0.100	-0.097
Semester	2.788	3.100	0.312
German high school degree	0.955	0.850	-0.105
High school GPA	1.710	1.700	-0.010
Last math grade	1.819	2.156	0.337
Mother university degree	0.446	0.750	0.304**
Father university degree	0.561	0.650	0.089
Mother employed	0.862	1.000	0.138*
Father employed	0.909	0.850	-0.059
First-generation migrant	0.121	0.150	0.029
Second-generation migrant	0.106	0.300	0.194**
Patience	3.530	3.700	0.170
Risk aversion	3.311	2.900	-0.411
Conscientiousness	3.902	4.150	0.248
Neuroticism	3.212	3.275	0.063
Openness	3.500	3.575	0.075
Extraversion	3.303	3.575	0.272
Agreeableness	3.250	3.225	-0.025
Feedback aversion	2.278	1.900	-0.378*
Self efficacy	3.939	4.033	0.094
Motivation university studies	3.864	3.700	-0.164
Weekly hours invested in course	5.189	5.150	-0.039
Initial motivation	3.591	3.700	0.109
Points PQI	8.136	6.875	-1.261*
Observations	66	20	86

Notes: Balancing check between individuals who did or did not report their grades on all controls used in the analyses. Sample comprises of all participants who reported their grades and came from course I.

Table A4.21 : Balancing Check for all Control Variables for Course 2, by Grade Reporting

Variable	(1) Grade reported	(2) No grade reported	(3) (1) vs (2)
Female	0.515	0.444	-0.070
Age	19.777	19.486	-0.291
Business Administration	0.476	0.694	0.219**
Economics	0.311	0.167	-0.144*
Bus. and Econ. Education	0.068	0.000	-0.068
Minor Bus./Econ./Education	0.087	0.111	0.024
Other major	0.058	0.028	-0.030
Semester	1.350	1.389	0.039
German high school degree	0.903	0.833	-0.070
High school GPA	1.704	1.903	0.199*
Last math grade	1.801	1.920	0.119
Mother university degree	0.524	0.611	0.087
Father university degree	0.650	0.667	0.016
Mother employed	0.806	0.889	0.083
Father employed	0.845	0.889	0.044
First-generation migrant	0.146	0.306	0.160**
Second-generation migrant	0.262	0.111	-0.151*
Patience	3.602	3.398	-0.204
Risk aversion	3.005	2.639	-0.366**
Conscientiousness	3.830	3.542	-0.288*
Neuroticism	2.966	2.819	-0.147
Openness	3.388	3.556	0.167
Extraversion	3.306	3.681	0.375*
Agreeableness	3.252	3.042	-0.211
Feedback aversion	2.068	2.000	-0.068
Self efficacy	3.819	3.917	0.098
Motivation university studies	3.864	3.944	0.080
Weekly hours invested in course	7.318	6.875	-0.443
Initial motivation	3.515	3.444	-0.070
Points PQI	9.505	6.958	-2.547***
Observations	103	36	139

Notes: Balancing check between individuals who did or did not report their grades on all controls used in the analyses. Sample comprises of all participants who reported their grades and came from course II.

Table A4.22 : Descriptive Statistics from the Post-exam Questionnaire

Variable	(1) Control	(2) POSNEG	(3) NEGPOS	(4) (2) vs (1)	(5) (3) vs (1)	(6) (2) vs (3)
Difficulty Exam	2.262	2.062	2.172	-0.200	-0.089	-0.111
Difficulty PQ I	2.976	3.309	3.195	0.332*	0.219	0.113
Difficulty PQ II	3.643	3.975	3.885	0.332*	0.242	0.090
Usefulness PQ I	3.190	3.099	3.230	-0.092	0.039	-0.131
Usefulness PQ II	2.762	2.704	2.736	-0.058	-0.026	-0.032
Correct recall FB1 Yes/No	0.929	0.852	0.793	-0.077	-0.135*	0.059
Correct recall FB1 Elements		0.855	0.913	0.000	0.000	-0.058
Correct recall FB1 Ordering		0.881	0.937	0.000	0.000	-0.055
Usefulness Feedback PQ I	3.333	2.420	2.319	-0.913	-1.014	0.101
Feelings Feedback PQI	3.000	3.174	2.536	0.174	-0.464	0.638***
Correct recall FB2 Yes/No	0.929	0.889	0.862	-0.040	-0.067	0.027
Correct recall FB2 Points	1.000	0.944	0.933	-0.056	-0.067	0.011
Usefulness Feedback PQ II	2.897	2.556	2.587	-0.342	-0.311	-0.031
Observations	42	81	87	123	129	168

Notes: Descriptive statistics from the post-exam questionnaire (non-standardized), by treatment status.

Table A4.23 : Differences by Gender for all Outcomes

	Motivation	Performance		Beliefs			Study
		PQ II	Exam	PQ I	PQ II	Exam	hours
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.236	-0.303*	0.359*	0.028	-0.095	-0.064	0.025
	(0.188)	(0.181)	(0.194)	(0.134)	(0.259)	(0.095)	(0.152)
Female=1	-0.322	-0.578**	-0.105	-0.363*	-0.534**	-0.396***	0.193
	(0.278)	(0.247)	(0.273)	(0.214)	(0.249)	(0.140)	(0.221)
POSNEG=1 × Female=1	0.102	0.376	-0.456*	-0.096	0.296	0.220	-0.122
	(0.250)	(0.255)	(0.253)	(0.181)	(0.322)	(0.147)	(0.217)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment performance	✓	✓	✓	✓	✓	✓	✓
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.347	0.335	0.538	0.678	0.120	0.758	0.575

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with being female. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, pre-treatment performance. Finally, all columns (besides column 5) include pre-treatment outcomes. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.24 : Differences by Initial Performance for all Outcomes

	Motivation	Performance		Beliefs			Study
		PQ II	Exam	PQ I	PQ II	Exam	hours
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.287**	-0.094	0.123	-0.025	0.066	0.058	-0.042
	(0.132)	(0.128)	(0.133)	(0.090)	(0.143)	(0.075)	(0.102)
Points PQ I (stand.)	-0.064	0.445***	0.690***	0.131	0.306***	0.048	-0.090
	(0.122)	(0.090)	(0.116)	(0.090)	(0.109)	(0.072)	(0.095)
POSNEG=1 × Points PQ I (stand.)	-0.295**	0.087	-0.153	0.022	-0.353**	-0.015	0.077
	(0.136)	(0.127)	(0.129)	(0.094)	(0.168)	(0.079)	(0.117)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.367	0.329	0.531	0.678	0.144	0.755	0.575

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with initial performance in the first set of practice questions. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on. Finally, all columns (besides column 5) include pre-treatment outcomes. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table A4.25 : Differences by Socio-economic Background for all Outcomes

	Motivation	Performance		Beliefs			Study hours
		PQ II	Exam	PQ I	PQ II	Exam	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.380**	-0.012	0.009	0.094	0.145	0.189**	0.078
	(0.158)	(0.154)	(0.161)	(0.108)	(0.185)	(0.084)	(0.107)
No academic parent=1	0.196	0.017	-0.017	0.130	-0.005	0.238**	0.195
	(0.201)	(0.171)	(0.205)	(0.138)	(0.195)	(0.117)	(0.174)
POSNEG=1 × No academic parent=1	-0.295	-0.294	0.296	-0.411**	-0.270	-0.447**	-0.414*
	(0.282)	(0.288)	(0.281)	(0.195)	(0.307)	(0.183)	(0.241)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment performance	✓	✓	✓	✓	✓	✓	✓
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.352	0.334	0.533	0.687	0.122	0.765	0.582

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the respective exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with socio-economic background as measured by having none or at least one parent with a university degree. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on, pre-treatment performance. Finally, all columns (besides column 5) include pre-treatment outcomes as well. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

5 Cognitive Skills Among Adults: An Impeding Factor for Gender Convergence?*

5.1 Introduction

Within economics, there has been much work on cognitive skills as a measure of educational success and as predictors of labour market outcomes. While the measurement of such skills among children and adolescents has been facilitated by the introduction of standardized tests within and across countries, cognitive skills among adults have been harder to study. This has in part been caused by the lack of internationally comparable measures of adult cognitive abilities. The OECD-administered PIAAC survey provides a solution to this missing data problem when studying adult skills. It provides measures of different cognitive skills as well as a rich set of background characteristics for adults aged 18-65 across 36 countries. This allows for a detailed assessment of these skills on a broad set of individuals from different backgrounds.

Getting a better understanding of the cognitive skills of adults is important for several reasons. On the one hand, they can help us understand how early-life knowledge transmits into adulthood. On the other hand, they can be a missing piece to the puzzle of gender pay gaps among adults. Despite recent convergence of labour market outcomes of men and women in many advanced economies, there are still important areas of individuals' lives that contribute to the remaining gender differences, such as parenthood and differences in educational choices (Goldin, 2014; Bertrand, 2020). Gender differences in labour-market relevant skills might both be an additional factor contributing to these remaining differences, but might also operate through the channels identified in the literature that impede full gender convergence on the labour market (Adda et al., 2017).

In this paper, we exploit the richness of the PIAAC dataset to investigate gender differences in numeracy skills among adults. First, we document average gender gaps in numeracy skills across countries and highlight the importance of studying these gaps from a distributional perspective (section 5.3). Second, we study the relationship between numeracy skills and wages and highlight the relevance of numeracy skills in accounting for parts of wage levels and gaps beyond past wages (section 5.4). Lastly, we delve deeper into gender numeracy gaps themselves and look at heterogeneities of these gaps along the dimensions most commonly identified in the literature to have an impact on gender gaps in other labour market outcomes (section 5.5). To strengthen the argument of the resulting patterns in our cross-sectional

* This chapter is co-authored with Michele Battisti and Alexandra Fedorets. It is based on the paper 'Cognitive Skills Among Adults: An Impeding Factor for Gender Convergence?', mimeo.

setting, we selectively use the German extension of the PIAAC dataset where a short panel of skills for the surveyed adults is available.

This paper delivers five stylised facts, which are intended to open up areas of future research to enhance our understanding of the determinants and consequences of gender gaps in cognitive skills across adulthood. A first key result is that despite recent improvements in educational equality in younger cohorts, important differences in skills among adult men and women persist. In most countries we study, men have higher average numeracy skills than women. These gaps in numeracy skills are sizeable and cannot be fully explained by the most commonly used characteristics as well as educational and occupational decisions. On the contrary, average gaps tend to increase when accounting for educational decisions which is in line with the current literature highlighting how women are catching up especially in higher education. Secondly, we confirm that individual numeracy skills are important predictors of wages in a contemporaneous regression specification. In addition, we use the short German panel of the PIAAC dataset (where cognitive skills and wages were surveyed three years apart, in 2012 and 2015) to show that numeracy skills predict current wages even when one controls for past wages and a series of indicators of past decisions.

We derive two further stylised facts from the analysis of this dataset by investigating adult numeracy skills along the joint hourly wage distribution. Gender gaps in numeracy skills in favour of men are most pronounced at the two top deciles of the wage distribution. No numeracy gaps can be observed in the middle part of the distribution whereas they are visible but relatively smaller in the bottom half of the distribution. Additionally, the share of women is highest in the bottom wage decile and steadily decreases towards the top decile. Furthermore, analysing wage *returns* to numeracy skills along the wage distribution reveals interesting patterns: there is a large difference in returns to higher numeracy skills at the top of the wage distribution, again favouring men. Instead, returns do not differ in the middle part and are even higher for women in the lower half of the wage distribution. To derive these results, individuals are divided into two subgroups with numeracy skills either below or above the country-specific median. The resulting returns described here hence refer to the relative advantage of above-median numeracy skills as compared to below-median numeracy skills for men and women. Both observed patterns are in line with the so-called glass ceiling effect in wages, which describes how women face disadvantages in a series of labour-market outcomes especially in the highest-paid occupations.

Lastly, we investigate the extent to which these differences depend on parental status and the field of study of individuals. Gaps in numeracy skills at the top of the wage distribution are especially pronounced for parents and those who have completed their highest degree in a non-STEM field of study. Among individuals with children, women are underrepresented (overrepresented) at the top (bottom) of the wage distribution. Consistently, the difference in *returns* to higher numeracy skills is barely visible for individuals without children, and more evident for individuals with non-STEM fields of study compared to STEM. When decomposing the numeracy gaps along the distribution of numeracy skills using a RIF decomposition (a

Kitagawa–Oaxaca–Blinder framework adapted to distributional analyses, roughly speaking), children and fields of study also play a prominent role: they increasingly account for differences in numeracy skills when moving towards the top of the skill distribution. The importance of children in the unexplained part of the decomposition along the entire numeracy distribution links differences in numeracy skills to differential returns to having children for men and women. Instead, country- and cohort-specific institutional features cannot be reconciled very well with the observed patterns. This last stylised fact is well connected with the results from the literature trying to explain gender gaps in wages and working hours. The decisions individuals take about their fields of study and their fertility are hence not only important for explaining these immediately visible labour-market outcomes, but are also connected to our measure of skills of the respective individuals.

This paper contributes to three strands of the existing literature. First, there is a rather established literature studying gender differences in cognitive skills as a measure of education that has almost exclusively focused on children and adolescents. A sizeable part of this literature has looked at gender differences in math skills during compulsory schooling (e.g. Hyde et al. (2008) for the US and Contini et al. (2017) for Italy). Some of these papers document that gender gaps in favour of boys are especially prominent at the top of the respective skill distribution (Ellison and Swanson, 2010; Robinson and Lubienski, 2011; Contini et al., 2017) or even at both tails of the distribution with males being overrepresented at the top and the bottom of the math skill distribution in the US (Autor et al., 2020). We contribute to this literature by focusing on adults whose cognitive skills have largely remained understudied. There are two papers that use PIAAC data to study adult cognitive skills: Rebollo-Sanz and De la Rica (2020) focus on average differences in cognitive skills and Christl and Köppl-Turyna (2020) use Austrian data to document gender differences in skills, task and skill matching of workers, and the impact of these factors on the gender wage gap using quantile regressions. Our paper extends this distributional approach to the entire PIAAC sample and additionally delves into potential channels of the observed differences in numeracy skills and their wage returns.

A related and very large strand of literature provides ample evidence on gender differences in labour market outcomes, especially in the highest-paid occupations. Albrecht et al., 2003 played an important role in the diffusion of the concept of 'glass ceilings' in the context of gender differences in the labour market. They provide strong evidence that wage differences between men and women in Sweden in 1998 were larger at the top, and that this difference is not driven by characteristics they can control for. Similarly, Collischon, 2019 documents a large glass-ceiling effect in Germany and Arulampalam et al., 2007 provide evidence for glass ceiling effects (as well as 'sticky floor' effects) across eleven European countries. Blau and Kahn, 2017 show that the decrease in the gender wage gap over the last decades in the US was much slower at the top of the wage distribution, and identify gender differences in occupations and industries as an important dimension. They briefly discuss gender differences in numeracy skills, and the possible role they may play for selection into STEM occupations. However, the numeracy skills they refer to only measure math test scores in high school. Bertrand, 2018

reviews the recent literature exploring possible explanations of glass ceiling effects across countries. For example, Petrongolo and Ronchi, 2020 focus on the role of technological change and the increase in service jobs to explain labour market performances of men and women. This paper adds to this literature by shifting the focus to cognitive skills used on the labour market. Using previously unavailable, internationally comparable measures of numeracy skills paired with a rich set of background characteristics and other labour market outcomes, we provide a detailed distributional analysis of numeracy skills as a potential labour market outcome as well as in their relationship with other labour market outcomes.

Lastly, we contribute to the literature exploring wage returns to cognitive skills. Using the NLSY79 dataset, Bacolod and Blum, 2010 document an increase in the labour-market returns to cognitive skills and a corresponding decrease in the returns to motor skills. This may have benefited women, since women tend to sort into occupations requiring mostly cognitive skills, while men are more likely to be in jobs that emphasize motor skills. We are able to enrich their discussion by investigating the role of numeracy skills in isolation, and painting a more nuanced picture than the (largely positive) one they discuss.[1] Hanushek et al., 2015 investigate the returns to cognitive skills using PIAAC data, and document that returns to cognitive skills are on average insignificantly different for males and females in the group of countries they examine. Again, this stresses the importance of studying gender differences in skills in more depth, in order to reconcile evidence from the glass ceiling literature with the lack of differential *average* returns to skills between men and women.

5.2 Data

Our main data source is the Programme for the International Assessment of Adult Competencies (PIAAC), a survey of adult skills developed by the OECD. The PIAAC study delivers internationally comparable measures of adult competencies, similarly to what the PISA study does for 15-year-olds. The study focuses on the necessary cognitive skills for advancing at work and participating in society, with the main focus on numeracy[2] and literacy[3] skills. Additionally, some of the participating countries conducted tests on problem-solving in technology-rich

[1] This literature is in turn related to the work on skill depreciation. Edin and Gustavsson, 2008 use Swedish administrative data to document 'economically important' depreciation of general skills after work interruptions. Ortego-Marti, 2017 shows that the rate of skill depreciation varies across occupations and industries, and in particular hits those occupations that require more skills. Most recently, Dinerstein et al. (2022) document skill depreciation among teachers in Greece waiting for central assignment to a teaching position after finishing their university degree.

[2] Numeracy is defined as *the ability to access, use, interpret, and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life.* A numeracy test can include understanding of a time series on birth rates or understanding different temperature measurement scales.

[3] Literacy is defined as *the ability to understand, evaluate, use, and engage with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential.* For instance, a test on literacy includes a list of pre-school rules and a question on their comprehension.

environments.[4] The measurement of skills is based on assessments, i.e. tests including a series of questions for each particular domain. Each skill is measured on a 500-point scale.[5] In this paper, we mostly focus on numeracy skills since they have shown to be the most relevant in predicting wages (Hanushek et al., 2015) and are more comparable across countries.

In addition to the skill measures, PIAAC gathers information on a wide set of socio-economic characteristics and labour market covariates of the individuals. In particular, it includes educational attainment and field of study, current work status, occupation, wages and working time, labour market history etc. The richness of background information is an important advantage of this dataset that facilitates a thorough analysis of the factors influencing an individual's skills.[6] The survey was initially conducted in August 2011 to March 2012 in the OECD countries. In its second round (April 2014 to March 2015), PIAAC was carried out in nine additional countries, including new OECD members and some non-OECD countries.[7] In our study, we mostly use information on the 32 countries that provide information on skill levels and wages. Table A5.1 lists the countries entering our analysis and the sample sizes at our disposal. As we do not focus on international comparisons, we standardize test scores within each country to achieve a mean of zero and a variance equal to one (see also Data Appendix A5.1). The only exceptions to this are figures 5.1, A5.1, A5.2, and A5.3 where we also show international differences in skill levels. There, the respective skill measure is standardized across the entire country sample.

We acknowledge that the cross-sectional nature of our data source restricts the empirical analysis, as we cannot observe the accumulation process of skills within individuals. To the best of our knowledge, the only country that used the initial sampling of PIAAC for a longitudinal study was Germany. The resulting PIAAC-L dataset provides a unique setting to follow individuals and their skills over time, but has two main disadvantages: First, samples sizes are unfortunately too small to conduct thorough analyses of individuals characteristics. Second, the time span of the dataset covers only three years (from 2012 to 2015), which limits the variation in skill development we can observe. Nonetheless, we use this extension for some selected additional analyses that provide a few useful insights into skill accumulation, bearing in mind that these findings cannot necessarily be generalized for other countries of the international PIAAC study.

[4] Sample questions can be found at `https://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm`, last accessed on November 3, 2022.

[5] It is important to underline that the test scores measure crystallized intelligence in particular domains and cannot be interpreted as ability or the overall level of intelligence (Halpern, 2013). It is also important to keep in mind that – despite the overall goal of the PIAAC tests to reduce country or gender biases to a minimum – even the testing mode itself and particular questions may contain undetected bias (Schroeders et al., 2016). For example, Griselda (2022) shows that a substantial part of the gender gap in math performance in PISA standardized tests can be attributed to gender differences in responding to multiple choice questions.

[6] For a more detailed description of the variables used in the analysis see Data Appendix A5.1.

[7] The full list of participating countries and the survey schedule can be found at `https://www.oecd.org/skills/piaac/`.

5.3 Numeracy Skills of Men and Women

We begin by illustrating some cross-country evidence on gender skill gaps. Figure 5.1 is a scatter plot of standardized numeracy scores by country, with each data point referring to the average score of men (y-axis) and women (x-axis) in each country. We differentiate between (a) all individuals and (b) those with non-missing wages, since we often present joint analyses of wages and numeracy levels in later chapters. Both graphs contain a 45-degree line, where test scores would lie in case of equality between genders. Numeracy skills are standardized across the entire sample to reflect differences in numeracy levels between countries. In all countries of the PIAAC sample, men on average have higher numeracy scores than women such that the resulting data cloud lies entirely above the 45-degree line (figure 5.1a). In the subsample with non-missing wages (figure 5.1b), the picture is very similar with most data points being above the 45-degree line. For some countries, the respective data points are closer or on the 45-degree line, which probably reflects positive selection into the labour market. Comparing panel (a) and panel (b) of figure 5.1 is consistent with the view that lower labour market participation of women is associated with lower numeracy skills. The corresponding within-country gender gaps can be found in figure A5.1. For the purpose of comparison, figures A5.2 and A5.3 in the appendix depict equivalent data clouds for literacy and problem-solving. The figures reveal that gender disparity in literacy is much less pronounced and that there is a range of countries where women on average have higher literacy scores than men. Scores for problem-solving resemble the distribution of numeracy scores more closely, both for all adults and for employed individuals only.

To get a better sense of the magnitude of these average gender gaps in numeracy scores across countries, we also show them in regression form. Table 5.1 presents a cross-country regression of standardized numeracy scores on a female dummy and country fixed effects. Each column then adds a relevant control which will also be used in later parts of this paper to help explain gender gaps in numeracy scores. Overall, gender gaps in numeracy scores in favour of men are large and persistent. Furthermore, we can see from column 3 that controlling for an individual's education level actually increases the gender numeracy gap which is in line with the recent literature showing that women actually have surpassed men in terms of education levels. In turn, occupations and fields of study help explaining parts of the gender gap in numeracy scores (columns 4-6). Lastly, column 7 adds a variable indicating how much individuals use numeracy skills during their work (self-reported). Even though this variable reduces the gender numeracy gap, the gap is far from disappearing. This piece of evidence already hints at some individual characteristics and choices that might be relevant for the emergence and persistence of gender numeracy gaps, but also shows that none of them will most likely be able to explain the entire gap.

Focusing on mean test scores only, may lead to an incomplete picture of men's advantage in numeracy. In fact, figure 5.2 (a) and figure 5.2 (b) show that the distributions of numeracy scores of men and women across countries substantially overlap, implying higher hetero-

geneity of test scores within gender than between men and women. Figure A5.4 depicts the gender-specific distributions of literacy and problem-solving scores, revealing that gender similarity in literacy is the highest, with an almost perfect overlap of the literacy score distributions of men and women.

In the following analyses, we often group numeracy skills into two categories: above the country-specific median and below. Table 5.2 shows descriptive statistics of men and women with "low numeracy" (defined as being below the country-specific median) and "high numeracy" (above the country-specific median).[8] We present descriptive statistics both for all survey participants (columns 1-4), and those with non-missing wages (columns 5-8). Among all participants, the proportion of men in the low-numeracy group is 45 percent, whereas it is nine percentage points higher in the high-numeracy group. The shares of younger age groups (20 to 29 and 30 to 44) are higher among the high-numeracy group, but the distribution of age groups does not show a distinctive gender pattern. The share of respondents living with their spouses is fairly evenly distributed between the two numeracy levels and genders.

A more distinctive pattern can be seen for respondents who have children. In the low-numeracy group, 68 percent of men and 79 percent of women have children, whereas in the high-numeracy group the share of respondents with children is six percentage points lower for men and eleven percentage points lower for women. This may partly be due to the fact that older respondents are over-represented in the low-numeracy group. As expected, lower education levels are more prevalent among the low-numeracy group: in both numeracy-level groups, more women than men have tertiary education. This is in line with the recent literature on women surpassing men on this dimension.

For many fields of study, we document relative gender parity, with some exceptions. In both numeracy groups, men study engineering, manufacturing and construction much more frequently than women, who in particular dominate in social sciences, business and law, as well as health and welfare. Studying social sciences, business and law, as well as science, mathematics and computing is much more frequently associated with higher numeracy levels for both genders. Among men and women with lower numeracy scores, STEM fields of study are less frequent than among the high-numeracy group.[9] Furthermore, women choose a STEM field of study less often in both numeracy groups. As for occupations, the most frequent ones in the low-numeracy group are craft and related trades workers for men and service and sales workers for women. In the high-numeracy group, men and women belong most frequently to the occupation group of professionals.

[8] The underlying distribution uses all individuals with non-missing numeracy scores without any further restrictions. The same classification is used in all analyses using numeracy above or below the median, independently of other restrictions applied to the respective samples.

[9] The fields 'Science, mathematics and computing' and 'Engineering, manufaturing and construction' are classified as STEM.

The share of employed increases from the low- to the high-numeracy group, both for men (from 73 to 83 percent) and especially for women (from 57 to 72 percent). Among these, the share of full-time workers is around 90 percent for men in both skill groups, whereas it is much lower for women (72 percent in the low-numeracy group and 77 percent in the high-numeracy group). The average wage grows with numeracy levels for both genders, although the raw wage gap is much higher in the high-numeracy group. Looking at wage percentiles, we observe that the pay gap widens with numeracy levels and is especially high among high-numeracy top earners. In general, these descriptive statistics demonstrate that numeracy levels are closely linked to labour market activity, wages, the probability of having children, and some fields of study and occupations. Columns 5-8 of table 5.2 show that individuals with non-missing wages are much more likely to have high levels of education and have a slightly different age structure, especially among those with below-median numeracy skills.

This section shows that gender disparities in numeracy skills among adults are pervasive and cannot be fully explained by differences in standard labour market characteristics. Beyond these average numeracy gaps, we are particularly interested in distributional analyses due to the large overlap of men's and women's numeracy skill distributions. Our focus on numeracy skills is rooted in their relevance in the labour market. In the following section, we discuss wage returns to numeracy skills.

5.4 Numeracy and Wages

5.4.1 Average Returns to Skills

In the following, we explore how numeracy skills are related to wages, and which insights skill gaps can provide on the formation of the gender pay gap. Because of the cross-sectional nature of the data, we observe skills and wages simultaneously.[10] Their relationship could hence go in both directions: individuals with higher skills tend to have better-paying jobs, but at the same time a better-paying job most likely requires more practice of particular skills and thus helps to preserve skill levels. Table 5.3 illustrates how wages and test scores correlate on average. In line with the previous literature (Hanushek et al., 2015), we find that numeracy levels have a higher predictive power for wages than literacy or problem-solving skills, both when included individually as well as simultaneously. Additionally, table 5.3 includes interactions of the respective skill variables with a dummy for being female, showing that *average* returns to skills for men and women cannot be statistically distinguished for numeracy and literacy skills in the specifications where skills enter separately (Columns 4 and 6). This is not the case for problem-solving where returns to skills are higher for women (Column 8). This preserves

[10] Figure A5.5 in the Appendix shows the distributions of (log) hourly wages (adjusted by country-specific PPP) of men and women pooled across all countries. The two distributions substantially overlap, with an almost perfect overlap for the low tails of the distributions, and a widening gap at log wages of about 1.7. The latter can be explained by women's over-proportionate engagement in part-time work. This leads both to lower monthly earnings due to reduced working hours as well as to a penalty in hourly wage rates.

into the specification where all skills are included simultaneously (Column 9), although only a subset of countries assessed problem-solving skills. In this reduced sample, we can also observe a negative additional return to numeracy for women, which was not visible when including numeracy skills only.

5.4.2 Inter-Temporal Wage Patterns

Numeracy skill levels do not only explain current wages, but also matter for the evolution of wages over time. Table 5.4 exploits the panel structure of the German PIAAC-L data by showing the dependence of wages in 2015 on wages and numeracy skills in 2012. Column (1) shows the raw gender gap, column (2) confirms the existence of a gender pay gap in wages 2015 after controlling for age, education, field of study, occupational groups and full-time work in 2012. Column (3) reveals a positive dependence of wages in 2015 from wages three years before, and shows that the female dummy indicating the gender pay gap even loses statistical significance. This implies that past wages absorb factors driving the gender pay gap. Moreover, the interaction of past wages with the female dummy is small and negative, indicating that wage evolution over the observed three years was, on average, gender neutral. Column (4) shows the positive dependence of wages in 2015 from past numeracy levels. As expected, the correlation is smaller than with past wages, whereas the interaction with the female dummy is larger but still statistically insignificant. Most interestingly, column (5) shows that numeracy skills in 2012 have predictive power for wages in 2015 beyond what can be explained by past wages. This highlights the importance of looking at the emergence and development of numeracy skills beyond their importance for wage gaps. Column (6) additionally includes contemporaneous numeracy levels for men and women, which both remain insignificant in the presence of past numeracy and wages. However, the last column in particular shows that past numeracy for men is correlated with an additional wage premium, which is completely cancelled out for women. This implies that higher numeracy levels are associated with higher wage growth, but only for men.[11]

Another way to see the importance of numeracy skills for gender differences in wages, is to look at their contribution to explaining gender wage gaps. Figure A5.6 (A) depicts the result of a Kitagawa–Oaxaca–Blinder decomposition of the gender pay gap into the explained and unexplained parts without considering numeracy levels (left bar) and with additional controls for numeracy skills (right bar). It shows that, on average, numeracy levels contribute positively and substantially to the gender gap formation, whereas average returns to numeracy - as mentioned above - do not differ by gender and thus do not contribute to the gender pay gap. Figure A5.6 (B) performs the same decomposition by country and shows that considering numeracy levels increases the gender pay gap only in few countries (including Japan, Chile, and the Czech Republic), whereas they are a substantial factor for explaining gender pay gaps in most other countries. Again, returns to numeracy do not substantially contribute to the average gender pay gap in most countries.

[11] All presented specifications are robust to the exclusion of particular controls.

5.4.3 Distributions of Numeracy Skills and Wages

Average differences mask important heterogeneity in the gender-specific patterns of skills and wages. We therefore now turn to analysing the distributional aspects of numeracy gaps between women and men. Figure 5.3 plots the share of women as well as average numeracy scores for women and men along deciles of the joint hourly wage distribution in the pooled sample of all countries with wage information. We can see that the share of women monotonically decreases along the wage distribution: from about 60 percent in the first decile to less than 35 percent in the top decile (dotted line). The numeracy levels for both genders also show an almost perfect monotonicity, with average numeracy levels being lower for low-wage earners and higher for high-wage earners. However, this simple representation reveals an interesting gender-specific pattern: men (dashed line) have relatively higher numeracy levels at the bottom and especially at the top of the wage distribution, whereas numeracy levels around the median wage are virtually the same as those of women (solid line). Within the wage deciles, the distribution of skills is very compact (see the p10-p90 intervals in figure 5.3), pointing towards a close relationship between numeracy levels and wages, i.e. numeracy being a good predictor for the wage level.

But even the same numeracy levels can have differential returns along the wage distribution for men and women. In order to study this aspect, we perform a decomposition based on the re-centred influence function (RIF) as suggested by Firpo et al. (2009). For this purpose, we estimate the following regression specification:

$$\log(W_{ic}) = \alpha + \beta * \text{Female}_{ic} + \gamma NS_{ic}^{top50} + \delta NS_{ic}^{top50} * \text{Female}_{ic} + \mathbb{X}_{ic}\mu + e_c + \epsilon_{ic} \quad (5.1)$$

The dependent variable is the log hourly wage of an individual i living in country c. *Female* is a binary variable equal to one for female respondents and zero otherwise. NS_{ic}^{top50} indicates that a respondent's numeracy skill level is above the median in his/her country of residence (skill levels below the median are the base category).[12] We also include an interaction term of the female dummy and the numeracy level. Thus, $\hat{\gamma}$ captures the returns to having above-median numeracy skills for men, relative to those with below-median numeracy levels. $\hat{\delta}$ captures the additional returns from above-median numeracy levels for women, compared to men. In our basic specification, we only control for a set of dummies for age groups 30 to 44, 45 to 54 and 55 to 65 (with ages 20 to 29 as the reference category) and control for the country of residence. In further analyses presented in the appendix, we add more controls. We then estimate equation (5.1) at all nine decile borders. For illustrative purposes, we summarize the estimation results in figure 5.4, which depicts the relative returns to numeracy levels for men ($\hat{\gamma}$) and for women ($\hat{\gamma} + \hat{\delta}$). The figure also depicts the marginal effect for females ($\hat{\beta} + \hat{\delta}$) to represent the gender pay gap at the respective decile.

Figure 5.4 confirms the established empirical fact that the gender pay gap is increasing (i.e. worsening) from the bottom to the top of the wage distribution and is especially pronounced

[12] As in table 5.2, the median split is performed on all individuals with numeracy scores in the respective country.

at the top two deciles. The figure also reveals gender-specific patterns in returns to numeracy: below the median hourly wage, returns to high numeracy levels are slightly larger for women than for men. In the middle of the wage distribution, returns to above-median numeracy are roughly equal and for the top two deciles, returns to higher numeracy levels are much higher for men than for women. This is because returns to higher numeracy skills for women remain stable over the wage distribution and slightly decrease for very high numeracy levels, whereas men see an increase of their returns to above-median numeracy skills.[13]

Figure A5.8 provides the same graph resulting from an estimation of equation (5.1) with additional controls for (A) education levels and the field of study, (B) occupational categories, and (C) a full-time indicator. We observe that the general picture of gender-specific returns to numeracy among the top earners remains unchanged in all specifications. The gender-specific detachment of skill levels from wages of top earners suggests that the existence of a glass ceiling in wages is less related to skills themselves, but rather to other, unobservable factors related to skills (e.g. networks) that then in turn hinder skilled women from earning more. Furthermore, relatively higher returns to skills for women in the lower wage deciles may point at lock-in effects of women with high skill levels in the low-wage segments.

To show gender-specific patterns of full-time employment as well as the influence of children, in table A5.6, we add an indicator for having children as well as its interaction with the female dummy to the extended specification of equation (5.1). We observe positive returns to children for men that increase from the lower to the upper deciles of the wage distribution. For women, this positive return is entirely cancelled out. This points towards distinctively different wage settings for fathers and mothers, even after controlling for their education, occupation and working time schedule.

In this section, we have explored the connection of adult numeracy skills and their contemporaneous and past wages. Numeracy skills are strong predictors of wages, both in contemporaneous regressions as well as in the short German panel including past wages and indicators. The patterns we observe speak to the glass ceiling effect for wages, which has been observed and investigated in the literature. Women have both lower average numeracy skills as well as lower returns to higher numeracy skills at the top of the wage distribution. They are also highly underrepresented in this part of the distribution.

5.5 Possible Drivers of Gender Skill Differences

The evidence presented above reveals that women have a disadvantage due to both lower numeracy levels and lower wage returns to numeracy skills, especially at the top of the wage

[13] Figure A5.7 depicts the returns to numeracy for women relative to men by plotting the $\hat{\delta}$ stemming from a country-wise estimation of equation (5.1). With few exceptions, we observe a dominant pattern of returns to numeracy for women being higher for lower wages and decreasing with wage levels, so that above-median numeracy skills pay off less for women than for men among high-earners.

distribution. Hence, the question emerges whether these differences in numeracy levels as well as returns can be explained by women's current or past choices compared to men. In the following, we provide empirical evidence on some of the channels that may explain the differences in observed numeracy levels. These differences are largely an outcome of a series of past events and decisions of the respondents (which may, in turn, be highly selective on numeracy levels).

Depicting the average numeracy levels for men and women in five-year age groups (figure 5.5, (a) for all individuals, (b) for those with non-missing wages) is an illustrative point of departure. Within all age groups, mean numeracy scores are higher for men than for women, with the lowest gap for the youngest group. This pattern is especially striking among respondents with non-missing wages. Moreover, for women, numeracy scores peak at ages 25 to 30 and then decrease. Men's numeracy levels are also highest for ages 25 to 30, but then remain at about the same level for the age groups 30 to 35 and 35 to 40 before they decrease for older groups. Hence, gender differences in numeracy skills are not specific to any age range, but are instead present across the entire age distribution.

5.5.1 Short-Term Accumulation of Skills using Panel Data

As figure 5.5 relies on cross-sectional data from respondents of different ages, it cannot be interpreted as a development of skills over the life course. The documented pattern could be driven both by a cohort component (e.g. more engagement in science for women from younger cohorts) and a life cycle component (e.g. gender-specific skill depreciation with age). In particular, a dominant life cycle component may imply that a relatively more equal gender distribution of skills among the young can be eradicated over the course of their lives if there is no change in institutional settings for skill accumulation and depreciation.

Using the German panel dataset PIAAC-L, we are able to disentangle these two effects, albeit with a smaller national sample and a short time span. Rebollo-Sanz and De la Rica (2020) mention that age-related gender skill profiles are likely to depend on skill depreciation. With the PIAAC-L data, we can empirically test if skills depreciate over time. Figure 5.6 shows the changes in skill levels for both genders by age groups. Among the youngest age group in 2012 (20 to 29 years old), both men and women improve their numeracy skills over time (i.e. until 2015). Men's skill gains are larger than those of women, though not significantly. In the age group 30 to 45, both men and women improve their numeracy skills by about the same amount, the improvement is smaller than in the youngest age group though. In the age group 45 to 54, women have an insignificant skill loss, whereas men again improve their skills. Among the oldest group aged 55 to 65, we observe a skill loss for both men and women.

Figures A5.9 and A5.10 depict the average growth in numeracy scores by gender for further categories. Men across all numeracy quartiles except the highest one experience skill growth over time (figure A5.9 A). Strikingly, the largest growth of numeracy scores is detected in the lowest numeracy quartile. Moreover, for women this is the only decile where we document

growth. We also find similar numeracy growth for men and women without children and with one child, but not for respondents with two children where women seem to lose numeracy skills (figure A5.9 B). In line with numeracy growth being an attribute of young respondents, we find the largest numeracy growth for men in the lowest education category (figure A5.9 C). Among the different fields of study, we observe pronounced over-time numeracy growth among women in science, mathematics and computing (figure A5.10 A). Among 1-digit ISCO occupations, there are only three clear patterns (figure A5.10 B): men in elementary occupations gain numeracy skills over time whereas women in this category experience losses in numeracy skill. Furthermore, men in the category of professionals gain numeracy skills over time whereas women experience skill growth only among the group of sales and service workers. The employment status of individuals is also related to numeracy skill growth (figure A5.10 C): on average, men in both full-time and part-time employment gain numeracy skills, whereas women only seem to gain when out of employment. The latter might be driven by women in education. This evidence suggests that the phase of educational attainment is crucial to numeracy accumulation, but that on-the-job skill accumulation is only present among men.

Table 5.5 depicts the dependence of current numeracy levels from past numeracy skills for men and women. We observe a gender gap in current numeracy skills, even after controlling for past numeracy. Instead, the interaction term of past numeracy with the female dummy is insignificant. Adding a series of controls shows that the accumulation of numeracy barely changes when including the field of study (potentially, because it is a past decision), but is more affected by the inclusion of the current occupation and an indicator for full-time employment. Strikingly, the inclusion of the dummy variable of having children and its interaction with the female dummy implies that children affect the skill accumulation of women, but not of men (column 6). In this last specification, the coefficient on the female dummy decreases substantially in size and loses significance.

5.5.2 Heterogeneity by Parental Status

As suggested by table A5.6, parental status plays an important role for the gender-specific relationship between skills and wages. Figure 5.7 (A) plots both numeracy scores by gender and shares of women along the hourly wage distribution for individuals with and without children. The numeracy profiles for men and women with and without children respectively are almost overlapping, except at the very top of the wage distribution. In the highest wage deciles, there is a substantial gap between men's and women's average numeracy skills which is more pronounced for individuals with children. In fact, for men, there seems to be no difference in numeracy skills in the highest decile by parental status. Along the rest of the distribution, childless individuals tend to have higher numeracy skills than those with children. When turning to gender shares within each wage decile, a more differentiated picture emerges for individuals with and without children. Among women and men with children, women are vastly overrepresented in low paying jobs and highly underrepresented at the top of the wage

distribution. This pattern is similar for those without children, but it is less pronounced both at the top and at the bottom of the wage distribution.

Figure 5.7 (B) presents returns to numeracy skills and the gender pay gap along the wage distribution by parental status. For childless men and women (black lines), the gender gap is much smaller. Also, their returns to numeracy are constant, with minor exemptions for the first and last decile borders. For men with children (gray dashed line), we observe increasing returns to numeracy along wage deciles, whereas the skill returns of women with children (gray solid line) are slightly declining above the median. Together, this suggests a stronger favouritism with respect to skills of fathers compared to mothers.

The results from figure 5.7 may partly be driven by selectivity into parental status. In order to address this aspect, figure A5.11 (A) presents gender-specific numeracy profiles for men and women depending on the age when they had their first child. It illustrates that particularly women who had their first child at a young age, exhibit lower numeracy levels than men who had their first child at the same age, and also as women who had their first child later in life. Figure A5.11 (B) shows the residuals of regressing numeracy on education levels, which highlights an even higher discrepancy by gender and high selectivity on numeracy levels for fertility decisions.

5.5.3 Heterogeneity by Higher Degrees in STEM

Given that numeracy skills are especially required in STEM-related occupations, we provide heterogeneity analysis by STEM versus non-STEM fields of study (figure 5.8). Figure 5.8 (A) again shows numeracy skills along the wage distribution as well as shares of women. Women are highly underrepresented in STEM fields of study along the entire wage distribution. Instead, they are overrepresented at all deciles for non-STEM fields of study, except at the very top of the wage distribution.

Instead, for numeracy skills, the picture is more differentiated. Individuals with degrees in non-STEM fields of study generally have lower numeracy skills than those from STEM fields of study. Furthermore, in the non-STEM group there are barely any gender gaps in numeracy skills except at the top of the wage distribution where men outperform women. This pattern is reversed for STEM fields of study where women outperform men in the largest part of the distribution, i.e. from the fourth to the ninth decile. Instead, at the extremes of the wage distribution, men have higher average numeracy skills than women.

Figure 5.8 (B) depicts the gender pay gap and returns to skills along the joint hourly wage distribution. The gender pay gap for respondents educated in STEM-related fields of study remains constant over the wage distribution, whereas it increases with higher wages for non-STEM fields of study. Strikingly, we observe that women with education in STEM-related fields have substantially higher returns to their skills in lower wage deciles, pointing towards higher favouritism of women at the bottom of the wage distribution.

5.5.4 Country-Specific Institutions

The literature emphasizes that gender differences in cognitive abilities may already emerge in school (Kahn and Ginther, 2017). It is also recognized that overall conditions during the time of initial labour market entry are important for educational levels and the employment trajectory (Hampf et al., 2020; Arellano-Bover, 2022). Additionally, country norms regarding the role of women for child care may matter for labour market outcomes. Figure A5.12 plots country averages for the gender numeracy gap against the percentage of the ISSP-respondents who agree with the statement that 'mothers of children under school age should stay at home'.[14] No clear pattern can be observed, even though there is a weak positive relationship. It appears crucial to use the variance *within* countries to study these relationships. This has been recently discussed by Moriconi and Rodríguez-Planas, 2021, who investigate the role of gender norms on the motherhood employment gaps across 186 regions in 29 countries.

In the following, we reassess the gender gap in numeracy adding various controls for the influences and conditions that individuals in our sample faced when they were 15 years old, i.e. at an age where a young individual would start thinking about their future plans. Table 5.6 starts with a specification in column (1) that includes only the female dummy. This estimates a raw gender gap for all individuals in the sample who have non-missing numeracy scores and are no first-generation migrants since for the latter we cannot adequately control for the country-of-origin institutional conditions. The inclusion of age brackets as controls yields virtually the same result (column 2).

Specification (3) adds the demeaned country-specific unemployment rate in the year the respondent was 15 years old and its interaction with the female dummy, to control for the overall economic conditions in the country at that time. We observe a significant positive correlation of the unemployment rate with the numeracy level for women, but no large change in the gender numeracy gap which corresponds to the coefficient on the female dummy here. Column (4) adds demeaned labour force participation (LFP) rates for men and women at age 15, as well as their interactions with the female dummy. The LFPs and the unemployment rate as a variable mix show differential influence on men and women and their inclusion increases the gender numeracy gap by about one percentage point.

Specification (5) includes the demeaned version of a proxy for *females in science* that measures the aggregate share of female authors in astrophysics in a country during the years when the respondent was 14 to 16 years old.[15] Although it seems reasonable to include this proxy in such a context since it aims to depict the presence of female role models in science, it neither correlates significantly with numeracy levels in the presence of other controls, nor does it contribute to the gender numeracy gap. Column (6) adds demeaned parental educa-

[14] The International Social Survey Programme (ISSP) on 'Family and Changing Gender Roles' IV was mainly conducted in 2012 among almost 40 countries (ISSP Research Group, 2016).

[15] The choice of astronomy as a field relates to the availability of reliable data from specialized scientific libraries in STEM for many countries and a possibly long period of time. The data stems from `http://ads.harvard.edu`.

tion levels and their interaction with the female dummy to control for the influence of the family environment. In the presence of other controls, we observe that parental education strongly correlates with the numeracy levels, though the interaction terms show no significant difference by gender.

Finally, column (7) shows the results when using control variables on parental education and interactions of the country and the respondents' year of birth to control for all possible institutional factors that may vary by country and year. Compared to specification (1), we document a slight reduction in the numeracy gap by less than one percentage point. Overall, this evidence points towards a relatively small importance of the institutional factors for the formation of the gender numeracy gap.

5.5.5 Decomposition of Numeracy Gaps

In order to understand how the characteristics presented in table 5.2 contribute to the formation of the gender gap in numeracy, we perform an Oaxaca-type decomposition of an estimated unconditional quantile regression, as suggested by Firpo et al. (2009). In particular, we first estimate the following unconditional quantile regression:

$$NS_{ic} = \alpha + \mathbb{X}_{ic}\mu + e_c + \epsilon_{ic}, \tag{5.2}$$

where NS_{ic} denotes the standardized numeracy score of an individual i from country c. $\mathbb{X}_{ic}$ comprises of the individual-level characteristics from table 5.2: socio-demographics (four age groups and being a parent), educational groups (3 categories: primary, secondary, tertiary), fields of study, and current occupation. e_c is a country dummy that we include in the equation to control for differences in labour market institutions between countries. We estimate equation (5.2) separately for men and women, and then perform the Kitagawa–Oaxaca–Blinder-style decomposition of the gender differences in numeracy levels explained by the observed characteristics captured in $\mathbb{X}_{ic}$ and the unexplained part, given by the differences in the returns to these characteristics by gender (μ).

Figure 5.9 (A) presents the numeracy levels of men and women by numeracy decile (legend on the right axis), their difference, and its decomposition into the explained and unexplained parts (legend on the left axis). The figure documents that the numeracy gap grows slightly with increasing numeracy levels. The explained part actually contributes negatively to the differences in numeracy levels below the median, which implies that the observed characteristics should be associated with a smaller gender gap. Instead, at the top of the numeracy distribution, differences in observed characteristics explain a part of the gender numeracy gap. Overall, however, the unexplained part dominates, especially above the median.

Figure 5.9 (B) shows the percentage contribution of the broad categories of controls (socio-demographics, educational level, field of study, occupation, country dummy) to the numeracy gap formation. Figure A5.13 contains a detailed decomposition of the explained part of the

numeracy gap by socio-demographics, field of study, and occupation. Figure 5.9 (B) reveals that women tend to have educational levels and occupations that are associated with higher numeracy levels and, hence, these factors contribute negatively to the numeracy gap. At the same time, their choice of field of study contributes positively to the gap and is a factor whose importance increases with the numeracy level (with the largest contribution of Engineering, Manufacturing and Construction, see figure A5.13B). Overall, occupations explain a small proportion of the gap. Belonging to the group of Managers as well as to Craft and related workers increases the gender numeracy gap, whereas belonging to Professionals decreases it (see figure A5.13C). The country dummy that captures institutional differences, positively contributes to the gender gap in numeracy.

Figure 5.9 (C) presents the unexplained part of the gap associated with the same variable groups, whereas figure A5.14 provides details on the decomposition of the unexplained part by socio-demographics, field of study, and occupation. The largest contributors to the unexplained part of the gap (figure 5.9C) are returns to socio-demographics and educational levels (that both explain part of the gap). Instead, observed returns to the field of study should be associated with a smaller gap than the one observed in the data. When looking at the detailed decomposition (figure A5.14), we see that returns to having children and being in the occupation groups of Professionals and Craft and Related Trade Workers are all related to relatively lower numeracy levels for women, whereas studying Engineering, Manufacturing and Construction is related to higher numeracy levels for women. Overall, we conclude that the observed characteristics of women, and especially their low presence in STEM-related fields of study is associated with a higher gender numeracy gap. Women are over-proportionately present among professionals, which contributes negatively to the numeracy gap, but within this group they have lower numeracy levels. The presence of children is related to a substantial part of the numeracy gap.

These results highlight the importance of parental status and field of study for the main patterns we find. Women's disadvantage at the top of the wage distribution is especially pronounced among parents and those with a non-STEM field of study whereas the patterns for childless individuals are much weaker and even partially reversed for STEM fields of study. Despite these striking patterns, parenthood and field of study as well as country- and cohort-specific institutional factors and many other characteristics usually associated with gender gaps in labour market outcomes cannot fully account for the gender gaps in numeracy skills among adults.

5.6 Conclusion

This paper presents new evidence on gender differences in numeracy skills and their relation to wage gaps. We use direct skill measures from the PIAAC dataset to study this relationship, hereby focusing on numeracy skills since they have shown to be particularly predictive of wages. Using PIAAC gives us the advantage of an objective skill measure for adults, i.e. a

deeper insight at the contemporaneous levels of skills in contrast to past educational levels that are often used in the literature. At the same time, the contemporaneous nature of our skill measures means that they are both input factors for current and future skill levels and wages, as well as outcomes from past education, life events, as well as the institutional context accompanying skill accumulation.

We first study the relationship of numeracy levels with wages and document that, on average, higher skills translate into higher wages. This also applies when studying the longitudinal data from the short German PIAAC-L panel, where higher numeracy levels also correlate with higher wage growth. However, the described relationship of numeracy and wages is much weaker for women than for men. Looking at numeracy levels along the wage distribution reveals that men's numeracy levels exceed those of women at the bottom and especially at the top of the wage distribution. Using an unconditional quantile regression, we demonstrate that returns to numeracy are almost the same for women along the wage distribution, whereas they are increasing for men. We also observe these patterns when controlling for education, field of study, occupation, and children. This suggests that the absence of progressive returns to skills for women may be a factor impeding them from aspiring to and preserving higher numeracy levels in the long run.

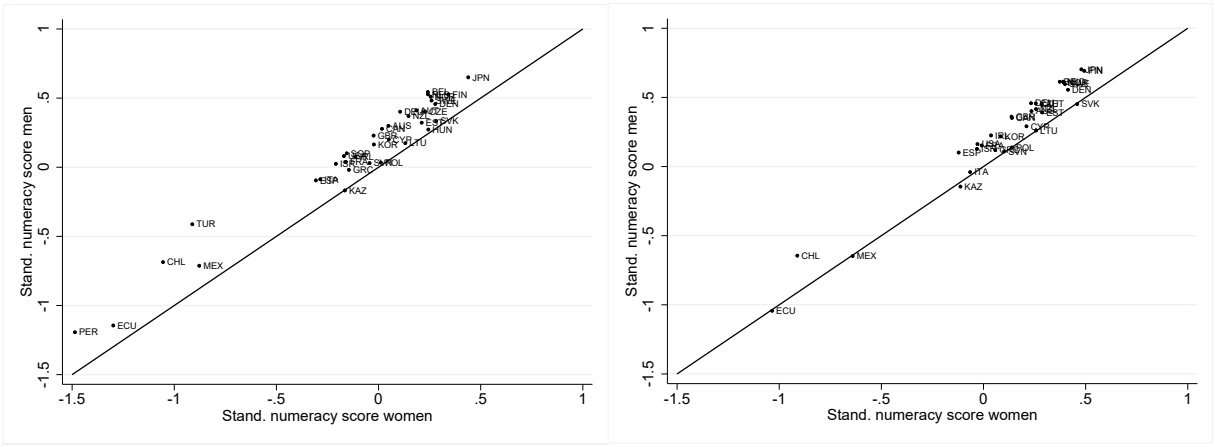
Indeed, the numeracy differences of men of women are smaller for younger cohorts and larger for the older ones, which may both be driven by different initial levels of numeracy at young ages and the influence of various events during the life course. Although we acknowledge that our main data source is unable to detect longitudinal changes due to its cross-sectional nature, we are able to empirically detect two driving factors of particular importance. First, we document that having children is associated with the numeracy levels of men and women. For childless men and women, the skill-cohort profiles and returns to skills along the wage distribution almost overlap, which is not true for mothers and fathers. Second, we detect that being educated in STEM-related fields of study is related to higher numeracy levels of both men and women. However, we also document that the returns to numeracy are particularly high for STEM-educated women in the low-wage sector. Concerning the country-level institutional factors, we do not find a strong impact on the gender numeracy gap. A decomposition of numeracy levels along its distribution confirms that the gender gap in numeracy largely depends on the field of study.

Overall, our results suggest that numeracy skills used on the labour market are an important driver for wages that is not stable over time and can accumulate or depreciate depending on labour market participation and family responsibilities. Therefore, our results support the importance of measures towards increasing numeracy levels of women by promoting STEM fields of study, but also underline that preserving numeracy levels and measures against its depreciation are of particular importance to women, especially for mothers. Our findings also point at undesirable patterns for returns to skills: favouritism of numeracy skills for women among low-wage earners and discrimination of their numeracy skills among top-earners.

Hidden factors like these may additionally discourage women from gaining and preserving higher numeracy levels.

Figures and Tables

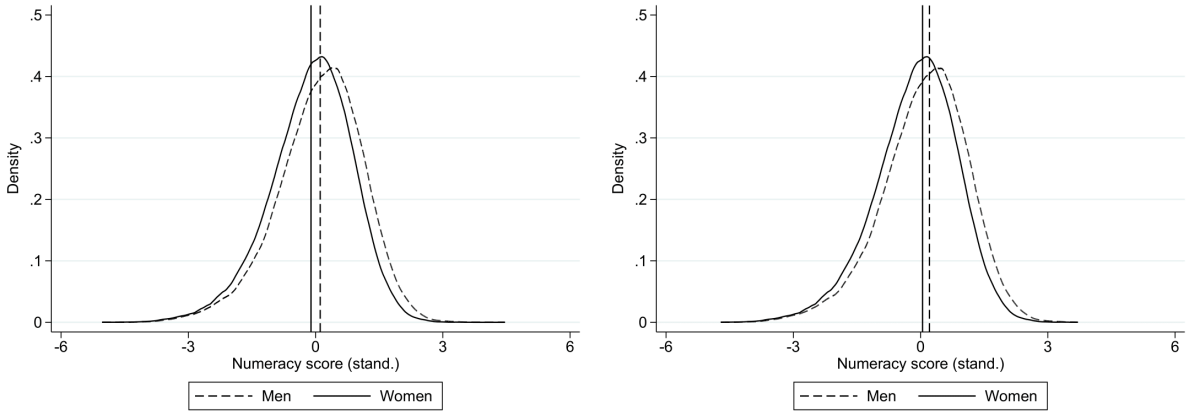
Figure 5.1 : Gender-Specific Numeracy Scores by Country



(a) All individuals (b) Non-missing wages

Notes: Standardized numeracy scores for men and women aged 20 to 65 by country. Standardization across all countries uses individuals' sampling probability. The graph additionally includes the 45-degree line to depict potential equality of test scores. Sample contains all individuals with non-missing numeracy scores (Panel a; 213,700 individuals) and non-missing wages (Panel b; 106,206 individuals). Data source: PIAAC international PUF 2012.

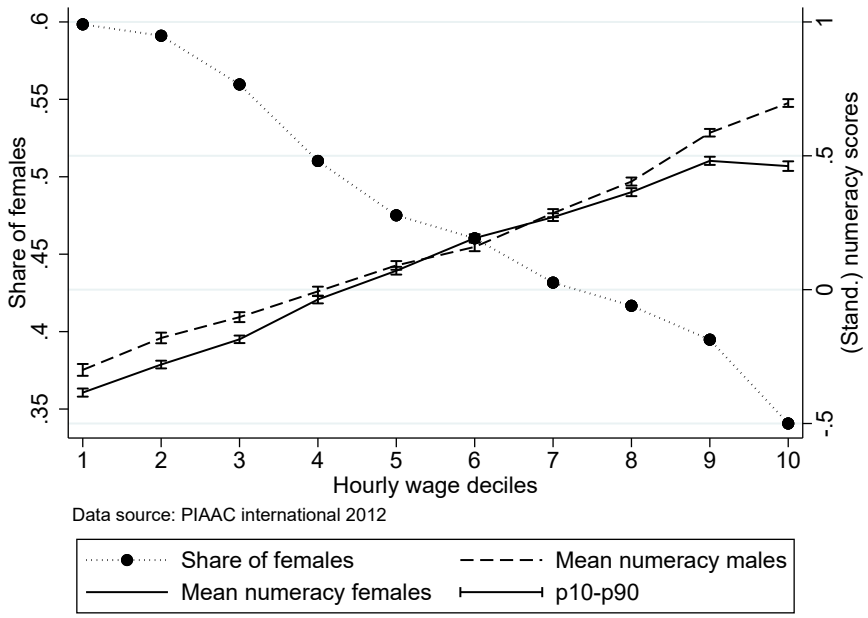
Figure 5.2 : Numeracy Score Distributions of Men and Women



(a) All individuals (b) Non-missing wages

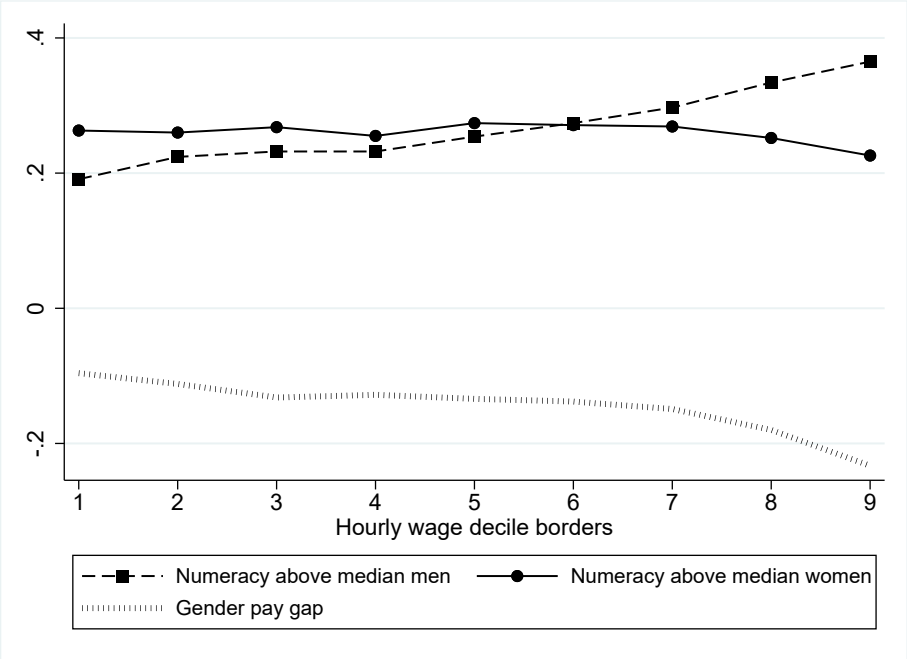
Notes: Standardized numeracy scores for men and women aged 20 to 65. Standardization by country uses individuals' sampling probability. Vertical lines represent the respective means for women and men. Sample contains all individuals with non-missing numeracy scores (a; 213,700 individuals) and non-missing wages (b; 106,206 individuals). Data source: PIAAC international PUF 2012.

Figure 5.3 : Numeracy Scores Along the Wage Distribution



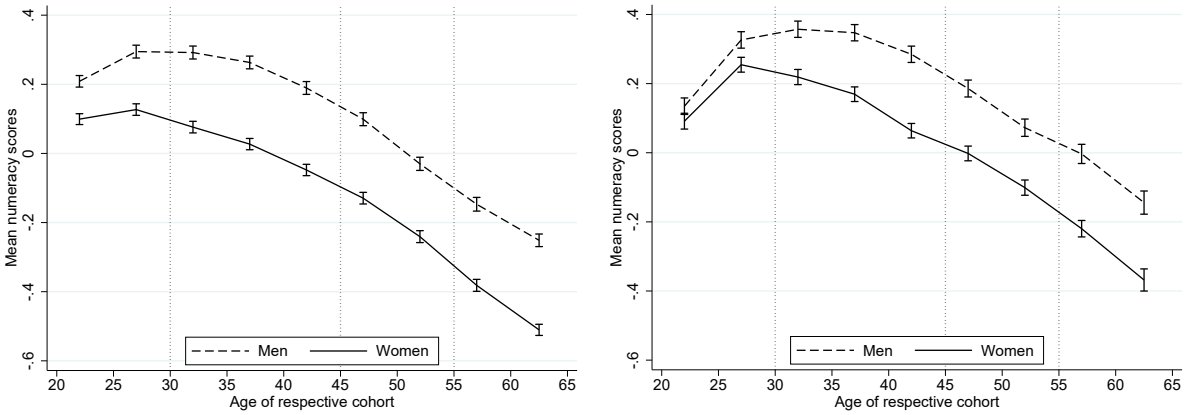
Notes: (Weighted) Shares of females within the respective deciles of hourly wages and standardized numeracy scores for men and women. Standardization by country uses individuals' sampling probability, deciles are calculated by country. Sample contains all individuals aged 20 to 65 with non-missing wages, i.e. 106,206 individuals. Data source: PIAAC international PUF 2012.

Figure 5.4 : Returns to Numeracy Levels, by Gender



Notes: Plot of the coefficients presented in equation 5.1 corresponding to unconditional quantile regressions without further controls (only age groups and country fixed effects) at each wage decile border. Graphs represent relative returns to numeracy levels for men ($\hat{\gamma}$, dashed lines with squares) and for women ($\hat{\gamma} + \hat{\delta}$, solid lines with circles) as described above. The dotted line plots the marginal effect for females ($\hat{\beta} + \hat{\delta}$) as described above. Corresponding coefficients can be found in table A5.2. Numeracy scores are standardized by country using individuals' sampling probability. Sample contains all individuals aged 20 to 65 with non-missing wages and numeracy scores, i.e. 106,206 individuals. Data source: PIAAC international PUF 2012.

Figure 5.5 : Numeracy Scores, by Age and Gender

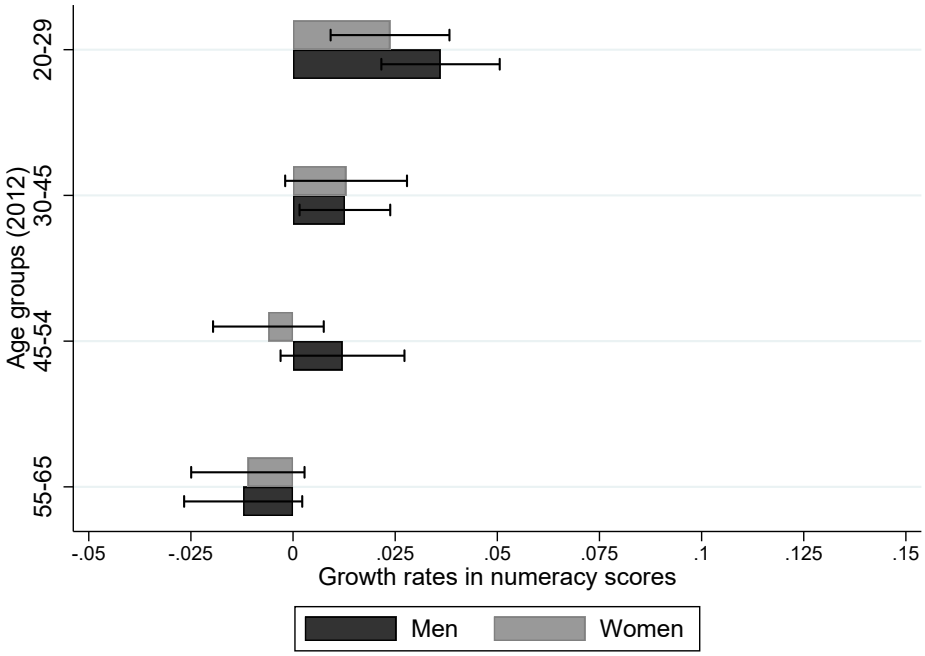


(a) All individuals

(b) Non-missing wages

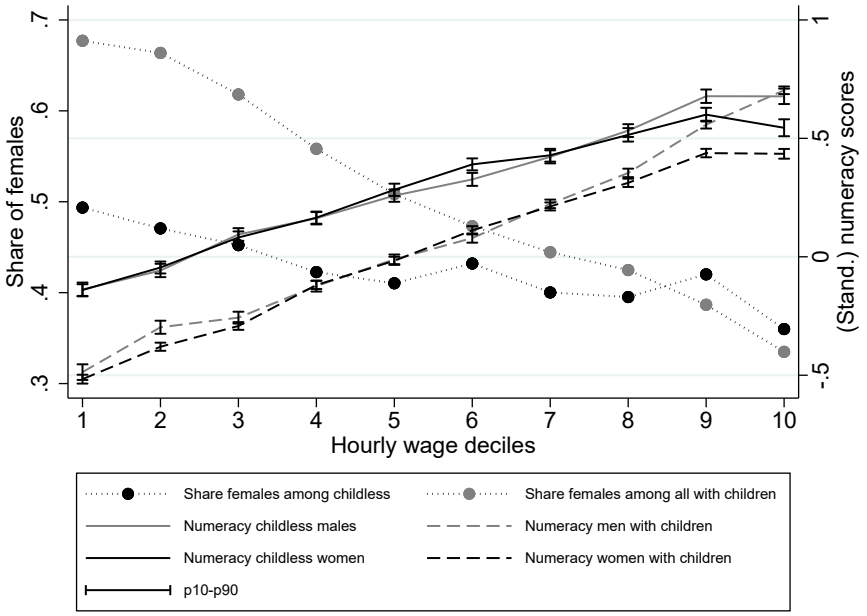
Notes: Mean standardized numeracy scores by age (in five-year intervals) for men and women aged 20 to 65. Confidence intervals for each data point are added, vertical lines represent cut-offs of age groups used in the regressions at ages 30, 45, and 55. Standardization by country uses individuals' sampling probability. Sample contains all individuals with non-missing numeracy scores and age (a; 213,700 individuals) and non-missing wages (b; 106,206 individuals). Data source: PIAAC international PUF 2012.

Figure 5.6 : Difference in Numeracy Scores between 2015 and 2012 for Women and Men, by Age (Germany only)

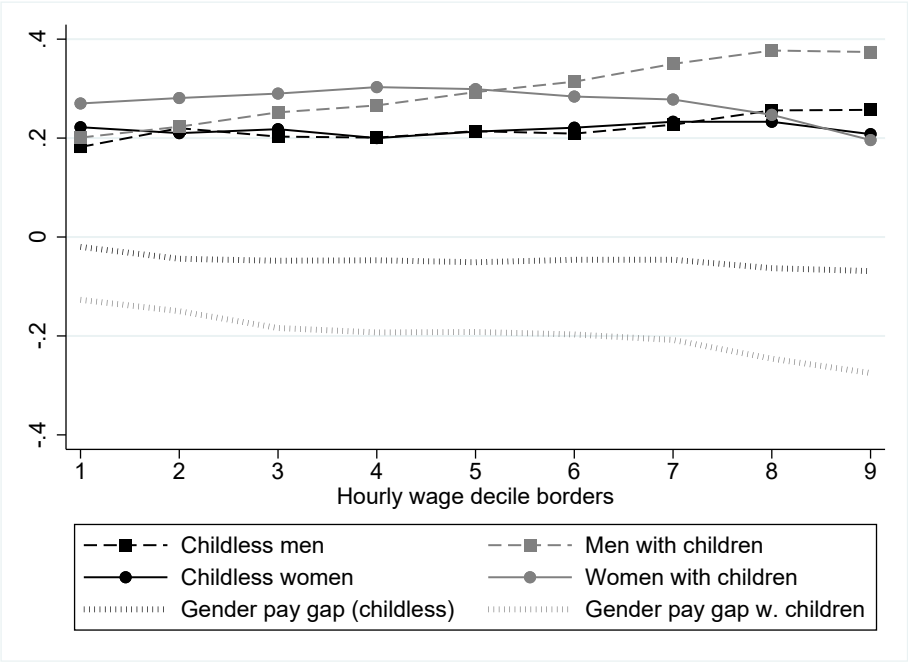


Notes: Growth rates in numeracy scores for men and women in Germany between 2015 and 2012 by age groups. Growth rates are calculated by dividing the difference between 2015 and 2012 numeracy scores by 2012 numeracy scores. Age groups refer to the age reported in 2012. Confidence intervals are added for bar. Sample contains all individuals aged 20 to 65 with non-missing numeracy scores in 2012 and 2015, and age in 2012 (2,961 observations). Data source: PIAAC-L German SUF 2015 and 2012.

Figure 5.7 : Parental Status, Numeracy Levels, and Wages



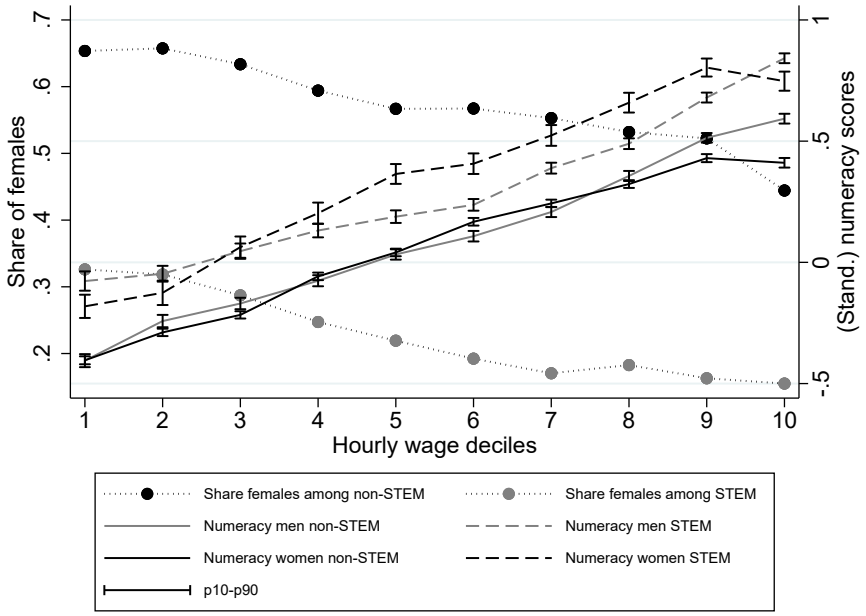
(A) Numeracy scores along the wage distribution, by gender and parental status



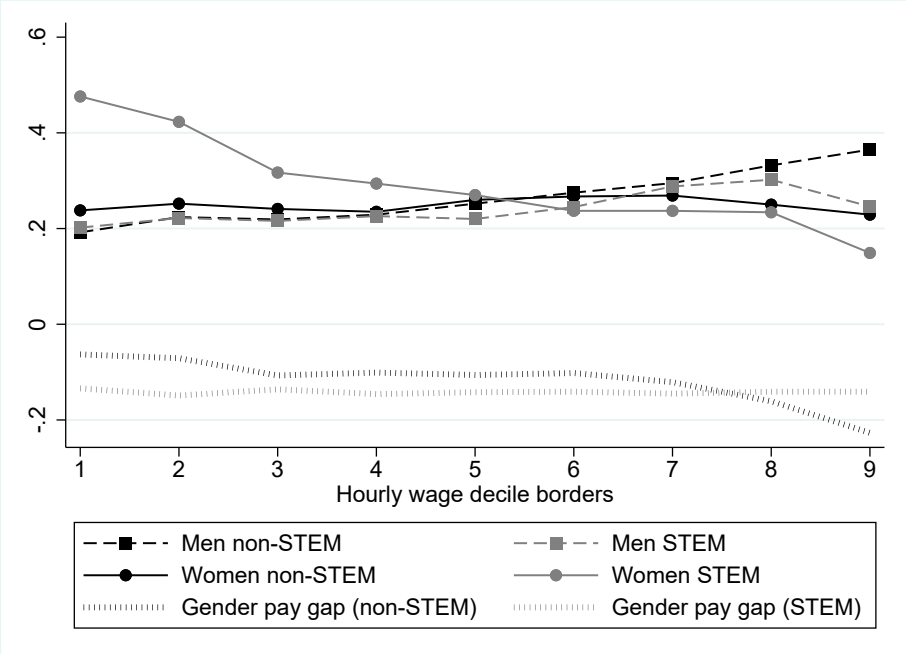
(B) Returns to skills, by gender and parental status

Notes: Panel (A): (Weighted) Shares of females within the respective deciles of hourly wages and standardized numeracy scores for men and women, by parental status. Standardization by country uses individuals' sampling probability, deciles are calculated by country. Sample contains all individuals aged 20 to 65 with non-missing wages. Panel (B): Relative returns to above-median numeracy levels for men and women by having children. The dotted lines plot the marginal effect for females. Corresponding coefficients can be found in tables A5.7 and A5.8. Numeracy scores are standardized by country using individuals' sampling probability. Sample contains all individuals aged 20 to 65 with non-missing wages, numeracy scores and information on children (106,134 individuals). Data source: PIAAC international PUF 2012.

Figure 5.8 : STEM v non-STEM Field of Study, Numeracy Levels, and Wages



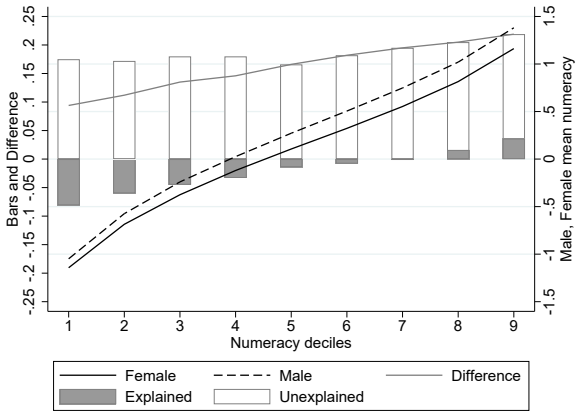
(A) Numeracy scores along the wage distribution, by gender and field of study



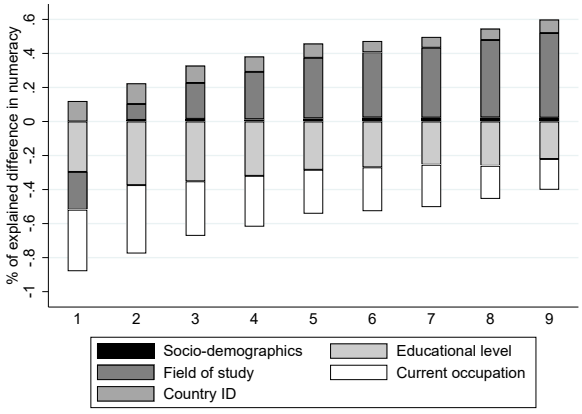
(B) Returns to skills, by gender and field of study

Notes: Panel (A): (Weighted) Shares of females within the respective deciles of hourly wages and standardized numeracy scores for men and women, by field of study. Standardization by country uses individuals' sampling probability, deciles are calculated by country. Sample contains all individuals aged 20 to 65 with non-missing wages. Panel (B): Relative returns to above-median numeracy levels for men and women by field of study. The dotted lines plot marginal effect for females respectively. Corresponding coefficients can be found in tables A5.9 and A5.10. Numeracy scores are standardized by country using individuals' sampling probability. Sample contains all individuals aged 20 to 65 with non-missing wages, numeracy scores and field of study (103,457 individuals). Data source: PIAAC international PUF 2012.

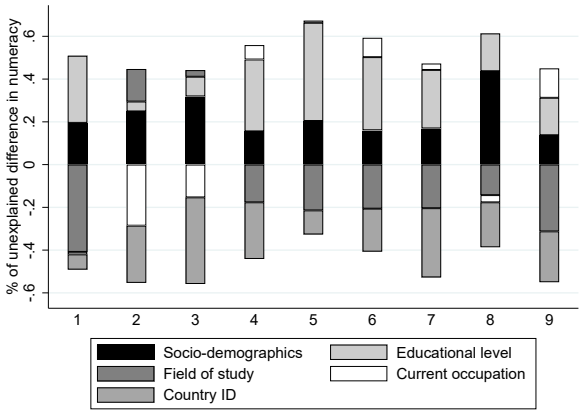
Figure 5.9 : Decomposition of the Factors contributing to Numeracy Scores



(A) Numeracy scores by decile: Explained vs. unexplained by observed characteristics



(B) Explained



(B) Unexplained

Notes: Kitagawa–Oaxaca–Blinder type decomposition of gender numeracy gaps by numeracy decile for employed individuals aged 20 to 65 (144,371 individuals) using the command *oaxaca_rif*. Explanatory variables used: age groups, children, education, field of study, occupation, and country dummies. Results look similar when just considering countries with earnings information or only individuals with non-missing wages for the explained part and differ slightly for the unexplained part (results available upon request). Data source: PIAAC international PUF 2012.

Table 5.1 : Gender Gaps in Numeracy Scores across Countries

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.217***	-0.211***	-0.228***	-0.212***	-0.203***	-0.201***	-0.197***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)
Age groups	No	Yes	Yes	Yes	Yes	Yes	Yes
Educational categories	No	No	Yes	Yes	Yes	Yes	Yes
Field of study	No	No	No	Yes	Yes	Yes	Yes
Occupational categories	No	No	No	No	Yes	Yes	Yes
Full-time indicator	No	No	No	No	No	Yes	Yes
Numeracy at work	No	No	No	No	No	No	Yes
Observations	213700	213700	213567	208000	144548	144022	115953
R^2	0.012	0.049	0.235	0.250	0.264	0.264	0.245

Notes: Dependent Variable: standardized numeracy scores. Least squares regression with country fixed effects, weighted by individual sampling probability. Estimation sample excludes all observations with missing values for the respective control variables. Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data source: PIAAC international PUF 2012.

Table 5.2 : Sample Description by Gender and Numeracy Level

	All participants				Non-missing wages			
	Low numeracy		High numeracy		Low numeracy		High numeracy	
	Men	Women	Men	Women	Men	Women	Men	Women
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share	0.45	0.55	0.54	0.46	0.48	0.52	0.55	0.45
Socio-demographics								
Aged 20-29	0.19	0.17	0.25	0.26	0.22	0.18	0.23	0.24
Aged 30-44	0.30	0.31	0.39	0.39	0.35	0.35	0.44	0.43
Aged 45-54	0.24	0.24	0.21	0.21	0.26	0.28	0.22	0.23
Aged 55-65	0.26	0.27	0.15	0.15	0.18	0.19	0.11	0.10
Live with spouse/partner	0.74	0.74	0.75	0.72	0.75	0.73	0.79	0.74
Has children	0.68	0.79	0.62	0.68	0.67	0.76	0.64	0.66
Education								
Lower secondary or less	0.36	0.35	0.12	0.10	0.29	0.22	0.09	0.06
Upper/post-secondary	0.49	0.44	0.45	0.40	0.54	0.49	0.44	0.37
Tertiary	0.15	0.21	0.43	0.50	0.17	0.28	0.47	0.58
Field of study								
General programmes	0.12	0.14	0.14	0.15	0.12	0.14	0.11	0.12
Teacher training and educ. science	0.02	0.06	0.03	0.10	0.02	0.08	0.04	0.12
Humanities, languages and arts	0.03	0.05	0.05	0.09	0.03	0.06	0.05	0.09
Social sciences, business and law	0.07	0.13	0.16	0.23	0.08	0.17	0.17	0.26
Science, mathematics and computing	0.04	0.03	0.11	0.08	0.04	0.04	0.11	0.08
Engineering, manufact. and constr.	0.27	0.05	0.31	0.07	0.32	0.05	0.33	0.07
Agriculture and veterinary	0.04	0.02	0.03	0.02	0.03	0.02	0.03	0.02
Health and welfare	0.02	0.10	0.03	0.11	0.02	0.15	0.03	0.14
Services	0.06	0.09	0.04	0.06	0.07	0.10	0.05	0.06
Missing (lower secondary education or less)	0.35	0.34	0.11	0.09	0.27	0.20	0.08	0.05
Field of study STEM	0.31	0.08	0.42	0.15	0.36	0.09	0.45	0.15
Occupation								
Armed forces occupations	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00
Managers	0.06	0.04	0.13	0.08	0.04	0.03	0.11	0.07
Professionals	0.07	0.14	0.23	0.31	0.07	0.15	0.23	0.32
Technicians and associate professionals	0.10	0.13	0.17	0.19	0.11	0.14	0.18	0.19
Clerical support workers	0.05	0.12	0.06	0.14	0.06	0.14	0.07	0.15
Service and sales workers	0.14	0.31	0.10	0.19	0.13	0.31	0.10	0.17
Skilled agric., forestry & fishery workers	0.06	0.02	0.03	0.01	0.03	0.01	0.01	0.00
Craft and related trades workers	0.24	0.04	0.15	0.02	0.24	0.03	0.14	0.02
Plant and machine operators/assemblers	0.17	0.04	0.09	0.02	0.19	0.04	0.09	0.02
Elementary occupations	0.11	0.15	0.04	0.05	0.12	0.15	0.05	0.05
Labor Market								
Share employed	0.73	0.57	0.83	0.72	1.00	1.00	1.00	1.00
Share full-time employed	0.90	0.72	0.91	0.77	0.92	0.73	0.93	0.79
Average wage	2.44	2.37	2.71	2.57	2.44	2.37	2.71	2.57
Wage p10	1.53	1.48	1.78	1.68	1.53	1.48	1.78	1.68
Wage p90	3.20	3.13	3.52	3.35	3.20	3.13	3.52	3.35
Observations	45,218	61,624	54,215	52,643	21,439	26,046	30,406	28,315
Observations (numeracy groups)	106,842		106,858		47,485		58,721	
Observations (availability wages)	213,700				106,206			

Notes: Descriptive statistics for men and women aged 20 to 65 by numeracy levels above or below the median, using sampling weights. Numeracy medians calculated by country. Field of study STEM refers to categories "Science, mathematics and computing" and "Engineering, manufacturing and construction". Sample contains all individuals aged 20 to 65 with non-missing numeracy scores. Data source: PIAAC international PUF 2012.

Table 5.3 : Returns to Skills: Regression of Log Hourly Wages on Skill Scores

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.149***	-0.180***	-0.166***	-0.166***	-0.176***	-0.176***	-0.168***	-0.170***	-0.159***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Numeracy (Num.)			0.068***	0.070***					0.058***
			(0.002)	(0.003)					(0.007)
Num. × Female				-0.004					-0.025**
				(0.004)					(0.009)
Literacy (Lit.)					0.057***	0.057***			0.018**
					(0.002)	(0.003)			(0.007)
Lit. × Female						-0.001			0.008
						(0.004)			(0.009)
Problem Solving (PS)							0.049***	0.041***	-0.006
							(0.002)	(0.003)	(0.005)
PS × Female								0.016***	0.027***
								(0.004)	(0.007)
Age groups	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Educational categories	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field of study	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupational categories	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Full-time indicator	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	106206	102576	102567	102567	102567	102567	74524	74524	74524
R^2	0.446	0.604	0.611	0.611	0.609	0.609	0.606	0.606	0.609

Notes: Dependent Variable: log trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if a continuous measure was not available. Skill measures are standardized at the country level using sampling probabilities. Least squares regression with country fixed effects, weighted by individual sampling probability. Dummies for education, field of study and occupation as well as a full-time indicator are included in columns 2-9. Baseline category for age groups is 20 to 29, the constant is omitted in the output. Sample contains all individuals aged 20 to 65 with non-missing data for wages as well as the respective controls. Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data source: PIAAC international PUF 2012.

Table 5.4 : Dependence of Wages in 2015 on Wages in 2012 and Numeracy Skills

	Outcome: Log Hourly Wages in 2015					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.089***	-0.080**	0.034	-0.055	-0.010	0.003
	(0.025)	(0.026)	(0.120)	(0.030)	(0.119)	(0.118)
Wages (2012)			0.514***		0.492***	0.490***
			(0.038)		(0.038)	(0.038)
Wages (2012) × Female			-0.018		0.004	0.001
			(0.043)		(0.043)	(0.043)
Numeracy (2012)				0.096***	0.058***	0.054**
				(0.020)	(0.017)	(0.019)
Numeracy (2012) × Female				-0.044	-0.034	-0.047
				(0.033)	(0.023)	(0.027)
Numeracy (2015)						0.008
						(0.017)
Numeracy (2015) × Female						0.022
						(0.026)
Age groups 2012	No	Yes	Yes	Yes	Yes	Yes
Educational categories 2012	No	Yes	Yes	Yes	Yes	Yes
Field of study 2012	No	Yes	Yes	Yes	Yes	Yes
Occupational categories 2012	No	Yes	Yes	Yes	Yes	Yes
Full-time indicator 2012	No	Yes	Yes	Yes	Yes	Yes
Observations	2006	1827	1734	1827	1734	1734
R^2	0.008	0.283	0.522	0.300	0.528	0.529

Notes: Dependent variable: log trimmed gross hourly wages in 2015. Least squares regression weighted by individuals' sampling probability. Dummies for education, field of study and occupation as well as a full-time indicator are included in columns 2-6. Baseline category for age groups is 20 to 29, the constant is omitted in the output. Sample contains individuals aged 20 to 65 and employed in 2012 and 2015 with non-missing data for wages, skill measures, gender, and all respective controls (in 2012). Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data source: PIAAC-L Germany 2015 and 2012.

Table 5.5 : Over-time Accumulation of Numeracy Skills.

	Outcome: Numeracy Scores in 2015						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.111**	-0.110**	-0.113***	-0.129***	-0.143**	-0.164***	-0.086
	(0.034)	(0.034)	(0.034)	(0.039)	(0.046)	(0.048)	(0.065)
Numeracy (2012)	0.750***	0.750***	0.716***	0.696***	0.644***	0.643***	0.644***
	(0.029)	(0.029)	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)
Numeracy (2012) × Female	-0.045	-0.045	-0.044	-0.034	0.007	0.011	0.002
	(0.040)	(0.040)	(0.039)	(0.039)	(0.042)	(0.043)	(0.043)
Children							-0.000
							(0.053)
Children × Female							-0.117
							(0.077)
Age groups 2012	No	Yes	Yes	Yes	Yes	Yes	Yes
Educational categories 2012	No	No	Yes	Yes	Yes	Yes	Yes
Field of study 2012	No	No	No	Yes	Yes	Yes	Yes
Occupational categories 2012	No	No	No	No	Yes	Yes	Yes
Full-time 2012	No	No	No	No	No	Yes	Yes
Observations	2961	2961	2960	2956	2353	2347	2347
R^2	0.502	0.502	0.507	0.514	0.487	0.488	0.489

Notes: Dependent variable: numeracy scores in 2015. Least squares regression weighted by individuals' sampling probability in the 2012-2015 sample. Sample contains individuals with non-missing numeracy scores in 2015 and 2012 as well as the respective controls, the constant is omitted in the output. Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data source: PIAAC-L German SUF 2015 and 2012.

Table 5.6 : Influence of Initial Labour Market Conditions on Numeracy Scores

	Outcome: Numeracy Scores (stand.)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.210***	-0.211***	-0.215***	-0.227***	-0.227***	0.220***	-0.202***
	(0.008)	(0.008)	(0.009)	(0.010)	(0.012)	(0.011)	(0.008)
Unemployment rate			0.077	0.376	0.160	0.060	
			(0.235)	(0.311)	(0.376)	(0.363)	
Unemployment rate × Female			0.703***	0.400	0.425	0.171	
			(0.197)	(0.283)	(0.368)	(0.360)	
Female LFP				0.229	0.386	0.431	
				(0.259)	(0.342)	(0.340)	
Female LFP × Female				0.237*	0.259	0.236	
				(0.115)	(0.162)	(0.161)	
Male LFP				0.001	-0.313	-0.714	
				(0.399)	(0.495)	(0.516)	
Male LFP × Female				-0.508*	-0.577	-0.778	
				(0.218)	(0.439)	(0.430)	
Females in science					0.072	0.078	
					(0.285)	(0.278)	
Females in science × Female					0.231	0.193	
					(0.306)	(0.301)	
Mother educ. intermediary						0.228***	0.196***
						(0.023)	(0.015)
Mother educ. intermediary × Female						0.045	0.064**
						(0.030)	(0.019)
Mother educ. high						0.418***	0.420***
						(0.027)	(0.018)
Mother educ. high × Female						0.052	0.029
						(0.036)	(0.024)
Father educ. intermediary						0.209***	0.219***
						(0.022)	(0.015)
Father educ. intermediary × Female						-0.019	-0.025
						(0.030)	(0.020)
Father educ. high						0.489***	0.467***
						(0.027)	(0.018)
Father educ. high × Female						-0.062	-0.061**
						(0.035)	(0.024)
Aged 30-44		-0.012	-0.016	0.032	0.054	0.028	0.010
		(0.030)	(0.032)	(0.037)	(0.042)	(0.038)	(0.049)
Aged 45-54		-0.070	0.017	0.246	0.308	0.505**	0.086
		(0.072)	(0.104)	(0.154)	(0.169)	(0.159)	(0.140)
Aged 55-65		-0.675***	-0.574***	0.077	0.123	0.508**	0.132
		(0.157)	(0.145)	(0.179)	(0.182)	(0.173)	(0.115)
Observations	88666	88666	78871	60282	36772	34477	82445
R^2	0.039	0.039	0.042	0.034	0.033	0.126	0.133

Notes: Dependent Variable: standardized numeracy scores. Least squares regression with country fixed effects as well as dummies for the year in which individuals were 15 years old, weighted by individual sampling probability. Column 7 additionally adds country-times-year15 fixed effects. All interacted variables are demeaned such that the coefficient on female can be interpreted as the resulting gender numeracy gap. Estimation sample excludes all observations with missing numeracy score and first generation migrants. Robust standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Results look similar when just considering countries with earnings information or only individuals with non-missing wages (results available upon request). Data source: PIAAC international PUF 2012, OECD (2020), ILO (2022), SAO/NASA (2022).

Appendix

A5.1 Data Appendix

General remarks

To make the procedures in this study comparable to related studies and to correct for some possible data issues, we perform some standard procedures on the data. As suggested by Hanushek et al. (2015) and Hampf et al. (2020), we remove the Russian Federation and Indonesia because the samples are not representative. In Russia, the Moscow region is entirely missing while in Indonesia, only the Jakarta region was sampled.

Control variables

The PIAAC survey offers a rich background questionnaire with many relevant information on individuals' personal lives and their labour market characteristics. The basic controls used in all presented regressions are a dummy for being female and age group dummies. The original gender variable provided in the PIAAC dataset is missing for one observation (out of the original 235,622) which is dropped in our entire analysis. Furthermore, a continuous measure for age is available for 181,005/235,622 observations. The missings come from Austria, Canada, Hungary, New Zealand, Singapore, and the US who only report age in five-year intervals from 16 to 65. In our study, we drop individuals aged 16-19 (18,865 observations) since we assume that most of these are still in education. Furthermore, in our regressions we only use age group dummies representing ages 20 to 29, 30 to 44, 45 to 54, and 55 to 65, the corresponding variable has no missing values.

The other socio-demographics presented in Table 5.2 refer to an individual living with their spouse or partner and their children. The former is taken as it is from the PIAAC dataset and has 31,080 missings of which only a small part comes from individuals aged 16-19 (564 observations). The remaining missings are within country and range from 3.19% in Singapore to 26.68% in Lithuania. The information on whether an individual has children is obtained from the top-coded version of a question on the number of children an individual has (top-coded at 4). The final variable then is 1 if the number if children is greater than 0 and is missing for 3,431 observations (from 0.05 % in Sweden to 13.87% in Cyprus).

An extended set of control variables includes information about individuals' education. The indicator for education levels is derived from a variable that distinguishes between six categories: Lower secondary or less (ISCED 1, 2, 3C short or less); Upper secondary (ISCED 3A-B, C long); Post-secondary, non tertiary (ISCED 4A-B-C); Tertiary: professional degree (ISCED 5B); Tertiary: bachelor degree (ISCED 5A); Tertiary: master/research degree (ISCED 5A/6); Tertiary: bachelor/master/reserach degree (ISCED 5A/6). We collapse all tertiary degrees into one indicator as well as the categories for upper and post-secondary education such that we obtain three categories for the education level of an individual (see table 5.2). This variable is missing for 3,108 observations; this number is composed of within-country missings ranging from 0.00% (Finland) to 13.83% (Cyprus).

A respondent's area of study in their highest qualification is reported in the categories presented in table 5.2. In this original version, the variable is missing for 63,842/235,662 observations (of which 12.158 among the 16-19 year olds that are dropped as described above). The remaining 51,684 missing values are mainly individuals with lower secondary education or less (43,039 observations), so we decided to add this as a category for field of study in order to not lose these observations in the regressions. The remaining 8,645 missing values are within-country missings (from 0.10% in Sweden to 42.95% in Israel).

Finally, we often control for an individual's occupation and working status. By doing so, we essentially restrict the sample to employed individuals since only those have non-missing information on their occupation (with the exception of 16 individuals aged 20 to 65) and their working hours. The categories used for the occupation refer to the 1-digit ISCO standard and are presented in table 5.2. The variable has 81,823 missings of which 13,439 come from individuals aged 16-19. The remaining 68,384 missing values almost entirely come from individuals who report not to be employed at the moment (unemployed or out of the labour force), only 2,684 employed individuals are missing this variable. Again, this comes from within-country missings ranging from 0.37% in Finland to 11.93% in Norway. Instead, the variables on employment status refer to an individual reporting to be employed as opposed to unemployed or out of the labour force as well as the reported working hours. The employment status of an individual is missing for 3,118 observations, most of which are aged above 19 and hence stay in our analysis (between 0.02% and 13.85% per country). Exploiting a question asking respondents to report their weekly working hours, we code a worker as employed full-time if the reported hours exceed 29 hours/week. The resulting variable is missing for 79,668 individuals of which 13,158 are 16-19 years old and will hence be excluded from our analysis. All non-missing values but one come from employed individuals and only 794 of the latter have missing information on their full-time status. Within-country missings for employed individuals range from 0.08% in Ireland to 2.4% in Israel.

Skill measures

Since in this study we are mainly interested in individual determinants and consequences of skill levels and gaps rather than international comparisons, we standardize the skill measures by country throughout the paper (if not specified otherwise). The three skill domains available in the PIAAC dataset (numeracy, literacy, problem-solving) are originally reported on a 500-point scale (OECD, 2016b). We standardize these measures to have mean 0 and standard deviation 1 within each country (using sampling weights). The exception to this are figures 5.1, A5.1, A5.2, and A5.3 where standardization is done across all countries using sampling weights in order to also show level differences in skills across countries. Throughout the analyses, we use the first plausible value of each skill measure (following Hanushek et al. (2015)).

Wages

Following Hanushek et al. (2015) and Hampf et al. (2020), among others, we perform a few important modifications to the available wage measures.

Not all countries provide continuous information on their respondents' wages. As can be seen in table A5.1, Peru does not provide any wage information at all whereas a number of countries only report the wage decile an individual is positioned in (Austria, Canada, Germany, Hungary, Singapore, Sweden, Turkey, and the US). For Germany, we are able to obtain a Scientific Use File that contains continuous wage measures. For Austria, Canada, Sweden, and the US, (as in Hanushek et al. (2015)) we are able to obtain country-specific information on each decile's median wage such that we can assign the decile median to each individual reported to be in the respective country-specific wage decile. This leaves us with four countries without wage information: Hungary, Peru, Singapore, and Turkey.

In the PIAAC questionnaire, individuals were asked about their preferred way of reporting their salary (*What is the easiest way for you to tell us your usual gross wage or salary for your current job?*). The response options ranged from the temporal frames *per hour* to *per year*, but there was also an option for piece rates. Depending on the answer to this question, individuals were forwarded to the question asking them to report the gross salary in their preferred way. Furthermore, if individuals were unsure or unwilling to report their salaries precisely, they were forwarded to a question where they got presented wage categories on the basis of their respective national earnings distribution in which they could place themselves as an estimate of their own salary. Similarly, bonuses and other additional payments were assessed. For self-employed individuals, only monthly earnings were asked.[1]

In this paper, we mainly focus on hourly wages due to their better comparability across individuals in different types of employment. The corresponding variables for hourly wages are reported both with and without bonuses for wage and salary earners, as well as ppp-adjusted and non ppp-adjusted (in US dollars). Wage deciles are available both for hourly earnings with and without bonuses. In order to obtain these measures of hourly wages from the reported earnings as described above, PIAAC performs a conversion of the given answers into both hourly and monthly earnings as described in OECD (2016a), chapter 20. The description here will focus on hourly earnings, details for monthly earnings can be found in OECD (2016a), chapter 20. As for hourly earnings, all salaries reported in categories other than *per hour* are converted into hourly salaries using the information about weekly hours worked from a previous question. For respondents who reported their earnings in intervals as described above, an imputation mechanism was developed. The imputation method would match each respondent with a "similar" respondent who reported earnings directly, where "similar" would be defined on the basis of highest education, skill level, age, and gender, among others. The precise earnings of this "similar" respondent were then used to impute

[1] See OECD (2016a) and `http://www.oecd.org/skills/piaac/BQ_MASTER.HTM`, last accessed November 03, 2022.

the respective earnings of the respondent who only reported wage intervals. This was done equivalently for bonuses/additional payments and monthly earnings. Furthermore, a variable indicating imputation of precise earnings was included (OECD, 2016a).

The readily available wage measures from the PIAAC dataset could in principle be used directly to conduct empirical analyses. Nonetheless, we perform some further adjustments to the wage data, following the procedure in Hanushek et al. (2015) and Hampf et al. (2020). As a first step, we assign decile medians as hourly earnings to further 21 observations, including a dummy indicating this procedure. In a second step, we trim 1 percent at the bottom and the top of the wage distribution in each country in order to reduce the possible influence of outliers. Finally, all wage measures are logged.

Sampling weights

To give the same weight to each country in pooled regressions, we standardize the sampling weights. The original variable *spfwt0* contains the final full sample weight provided by the OECD that makes sure each country is representative in a given dataset, both in size and regarding relevant demographic characteristics. Since we do not wish to represent different sizes of countries in our pooled regressions, we adjust this variable to sum up to 1 in a country instead of its effective size. These adjusted weights are then used in our regressions throughout (if not specified otherwise).

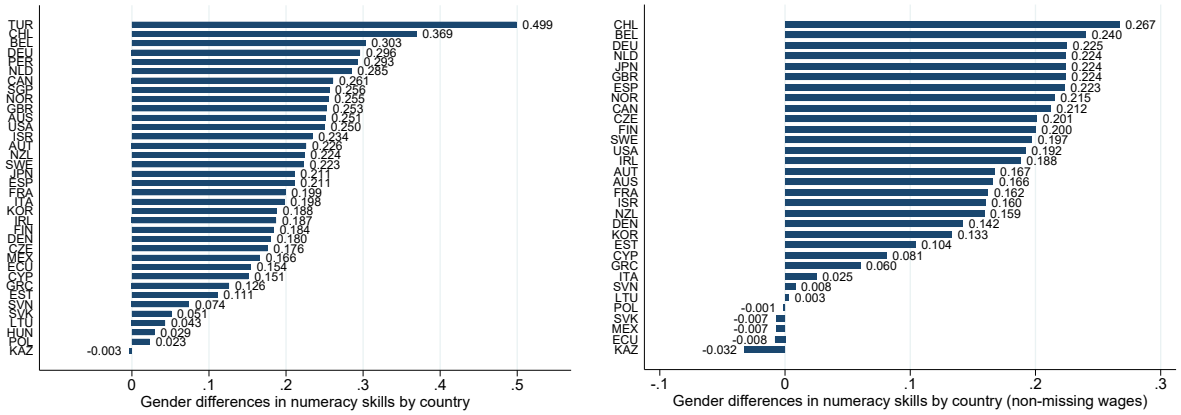
German panel dataset

As mentioned above, Germany assessed individual skills for the 2012 PIAAC sample again in 2015 to create a small panel dataset. Mostly, we apply the same corrections/transformations to the dataset as described in subsections A5.1 and A5.1. Hence, this subsection will focus on the differences only.

The resampling of German respondents took place from 2014 to 2016. In 2014, only household members aged 18 or above of the 2012 respondents were surveyed. In 2015, both the original 2012 respondents and their partners living in the household were surveyed in a similar way as in the original questionnaire in 2012, including a comparable skill assessment. The last sampling in 2016 again included household members aged 18 or above from the respective households. Since numeracy skills were only measured in a comparable way in 2015, we focus on the samples from 2012 and 2015 when using the German sample. Wages in 2015 are not available as a continuous measure but only in wage intervals. Hence, individuals are assigned the midpoint of this interval as their wage measure. Hanushek et al. (2015) show that this procedure in general provides very similar results to the use of continuous wages. In 2012, we have continuous measures for wages provided by PIAAC such that we decided to use the best available measure in each year.

A5.2 Appendix Tables and Figures

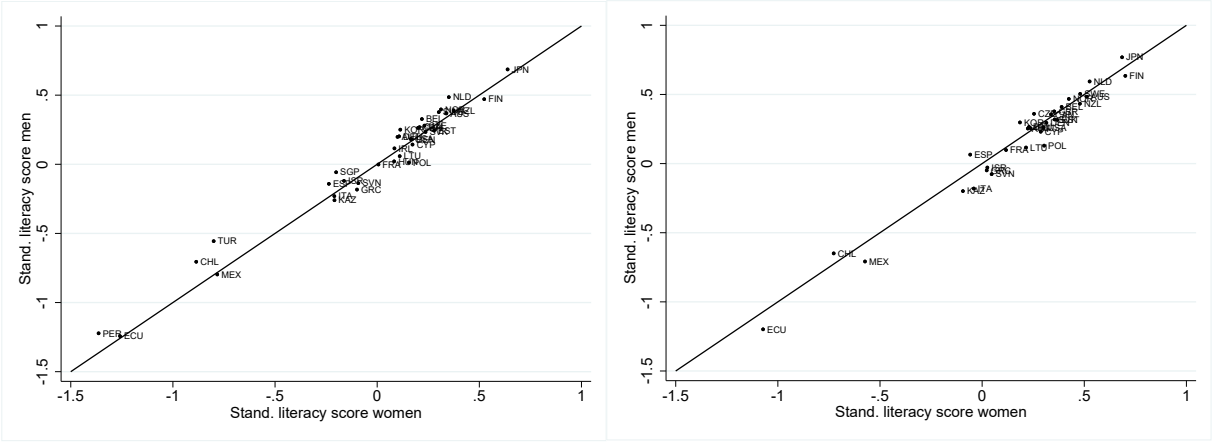
Figure A5.1 : Gender Gaps in Numeracy Scores by Country



(a) All individuals (b) Non-missing wages

Notes: Gender gaps in standardized numeracy scores for men and women aged 20 to 65 by country (all (a) or only those with non-missing wage (b)). Gender gaps represent coefficients for female of a regression of standardized numeracy scores on a female dummy, by country using sampling weights. Standardization across all countries uses individuals' sampling probability. Sample contains all individuals with non-missing numeracy scores (a; 213,700 individuals) and non-missing wages (b; 106,206 individuals). Data source: PIAAC international PUF 2012.

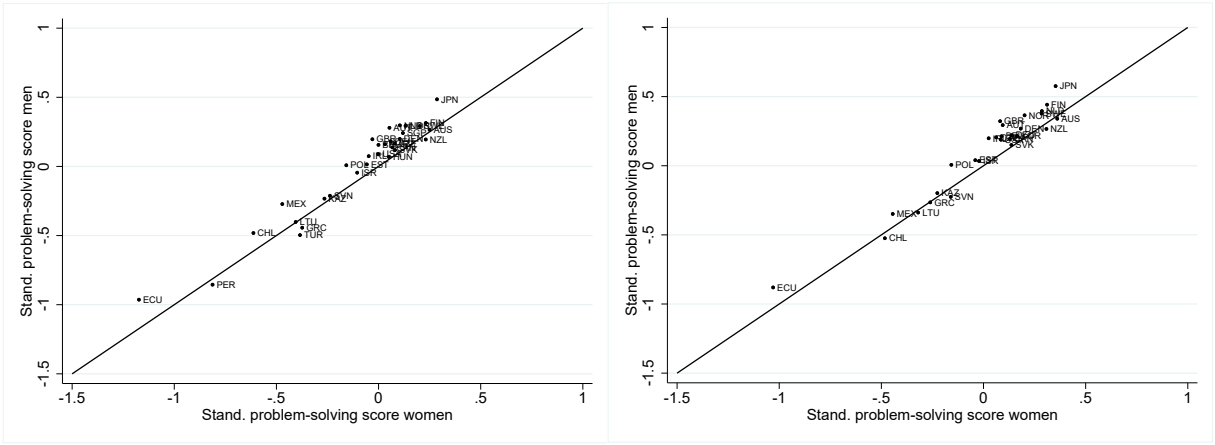
Figure A5.2 : Gender-specific Literacy Scores by Country



(a) All individuals (b) Non-missing wages

Notes: Standardized literacy scores for men and women aged 20 to 65 by country. Standardization across countries uses individuals' sampling probability. The graph additionally includes the 45-degree line to depict potential equality of test scores. Sample contains all individuals with non-missing literacy scores (a; 213,700 individuals) and non-missing wages (b; 106,206 individuals). Data source: PIAAC international PUF 2012.

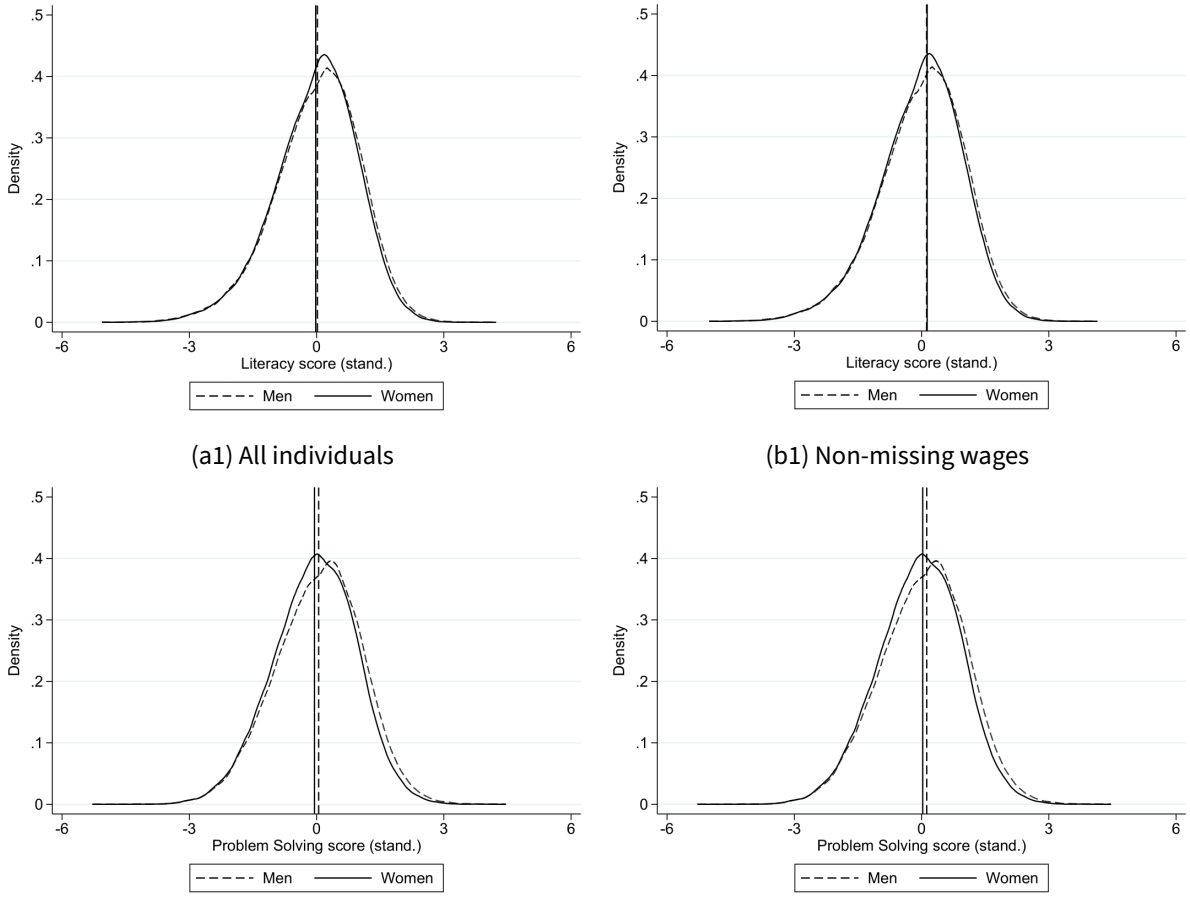
Figure A5.3 : Gender-specific Problem-solving Scores by Country



(a) All individuals (b) Non-missing wages

Notes: Standardized scores for problem solving in technology-rich environments for men and women aged 20 to 65 by country. Standardization across countries uses individuals' sampling probability. The graph additionally includes the 45-degree line to depict potential equality of test scores. Sample contains all individuals with non-missing problem-solving scores (a; 136,796 individuals) and non-missing wages (b; 77,301 individuals). Results look similar when just considering countries with earnings information (results available upon request). Data source: PIAAC international PUF 2012.

Figure A5.4 : Test Scores in Literacy and Problem-solving, by Gender



(a1) All individuals

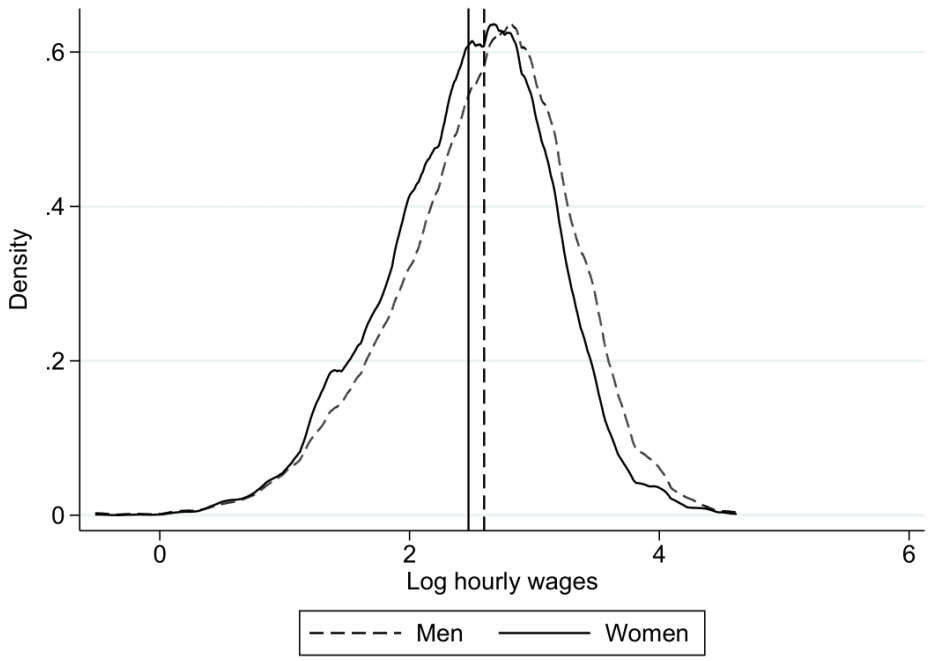
(b1) Non-missing wages

(a2) All individuals

(b2) Non-missing wages

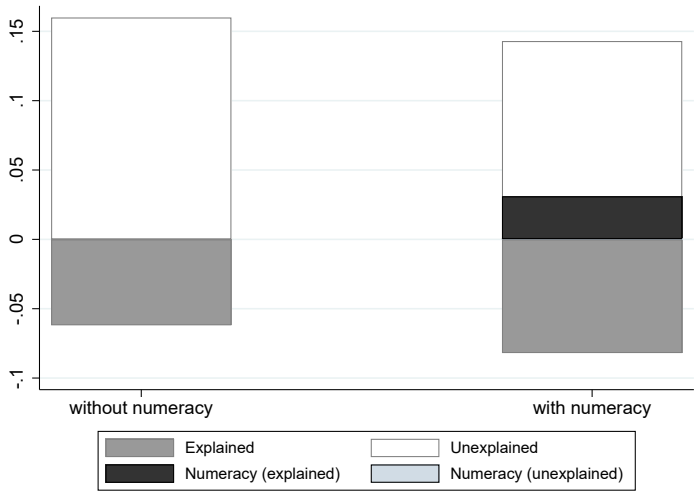
Notes: Standardized literacy and problem solving scores for men and women. Standardization by country uses individuals' sampling probability. Vertical lines represent the respective means for women and men. Sample contains all individuals with non-missing skill measures (a1: 213,700 individuals; a2: 136,796 individuals) and non-missing wages (b1: 106,206 individuals; b2: 77,301 individuals). Results look similar when just considering countries with earnings information (results available upon request). Data source: PIAAC international PUF 2012.

Figure A5.5 : Distribution of Gross Hourly Wages, by Gender

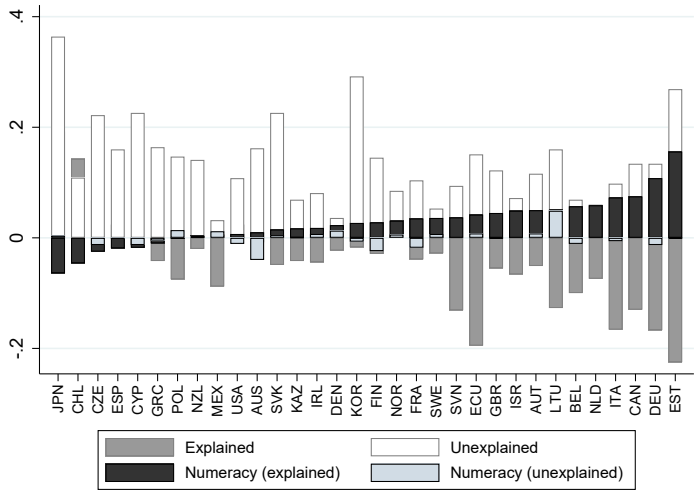


Notes: Log trimmed gross hourly wages (ppp-adjusted) for men and women. Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Vertical lines represent the respective means for women and men. Sample contains all individuals with wage information (106,206 individuals). Data source: PIAAC international PUF 2012.

Figure A5.6 : Role of Skills in Gender Gap Formation



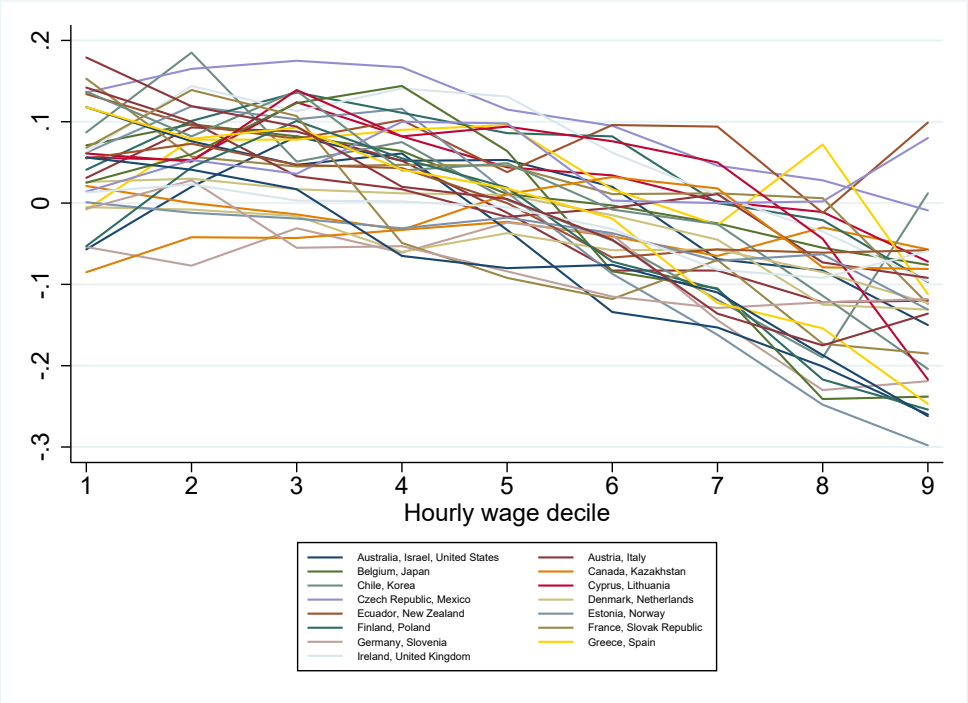
(A) On average



(B) By country

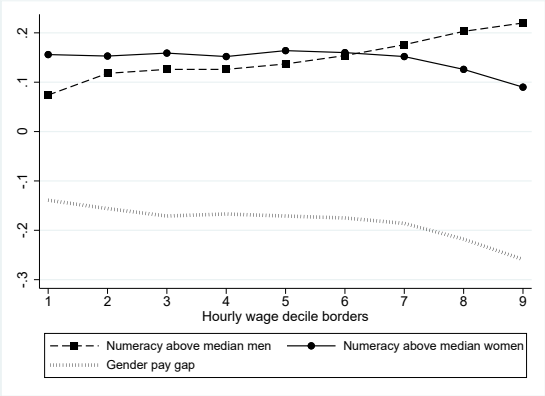
Notes: Kitagawa–Oaxaca–Blinder decomposition of gender gaps in hourly wages for employed individuals aged 20 to 65. Explanatory variables used: age groups, children, education, field of study, occupation, and country dummies. Numeracy scores are added as explanatory variables in second bar and in panel B. Sample contains all individuals with non-missing wages and numeracy scores (106,206 individuals). Data source: PIAAC international PUF 2012.

Figure A5.7 : Returns to Numeracy for Women relative to Men, by Country

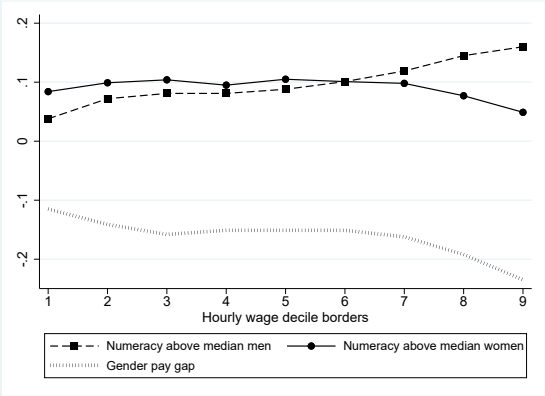


Notes: Plot of the coefficients presented in equation (5.1) corresponding to unconditional quantile regressions with full controls (age groups, education levels, field of study, occupation, full-time status, children, and children×female) at each wage decile border. Graphs represent returns to numeracy levels for women relative to men ($\hat{\delta}$) as described above. Sample contains all individuals with non-missing wages and numeracy scores as well as the respective controls (overall 102,506 individuals). Data source: PIAAC international PUF 2012.

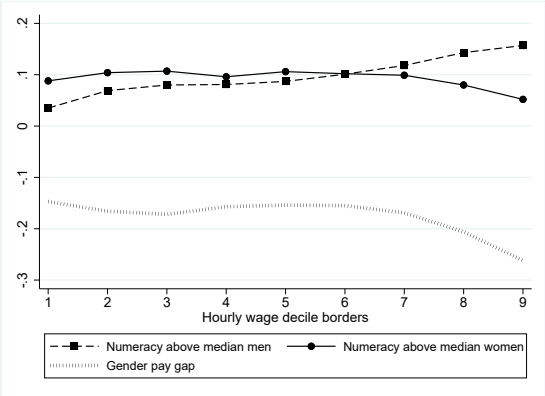
Figure A5.8 : Returns to Numeracy Levels with Additional Controls



(A) Additional controls: education level and field of study



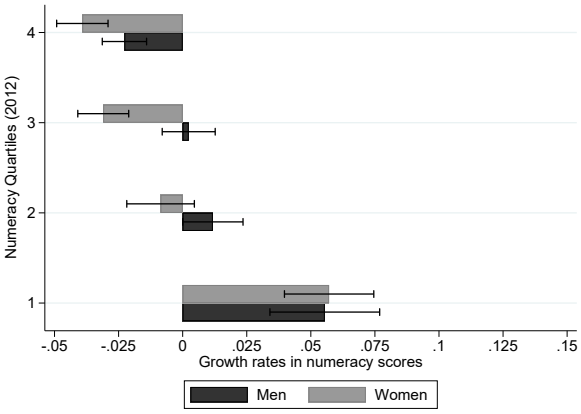
(B) Additional controls: occupation category



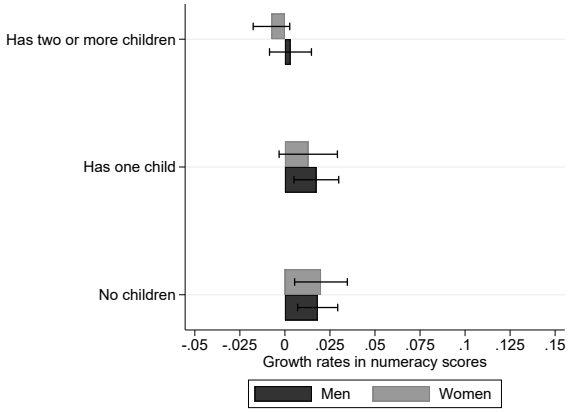
(C) Additional controls: full-time indicator

Notes: Plot of the coefficients presented in equation (5.1) corresponding to unconditional quantile regressions with further controls level of education, field of study, occupation, and a full-time indicator (in addition to age groups and country fixed effects) at each wage decile border. Level of education, field of study, and occupation are measured as presented in table 5.2. Field of study has an additional category for individuals with the lowest level of education and missing information on field of study. The full-time indicator takes on the value 1 if an individual is in full-time employment (more than 29 working hours per week) and 0 otherwise. For description of graphs see notes of figure 5.4. The corresponding tables can be found in table A5.3, table A5.4, and table A5.5. Numeracy scores are standardized by country using individuals' sampling probability. Sample contains all individuals aged 20 to 65 with and non-missing wages, numeracy scores, and the respective controls (A: 103,443 individuals; B: 102,602 individuals; C: 102,567 individuals). Data source: PIAAC international PUF 2012.

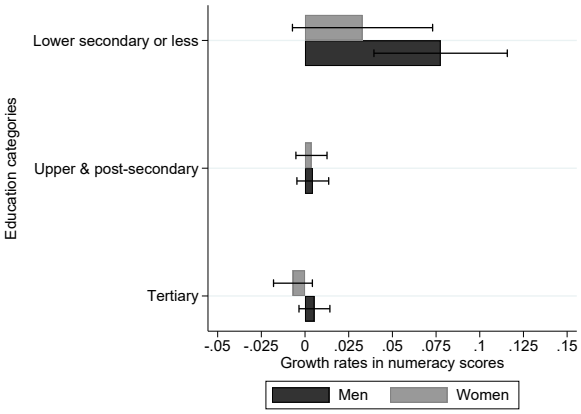
Figure A5.9 : Change in Numeracy Test Score between 2012 and 2015 (I)



(A) By quartile of numeracy score 2012



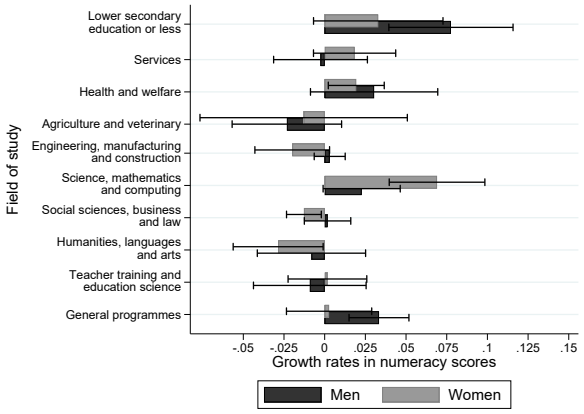
(B) By number of children 2012



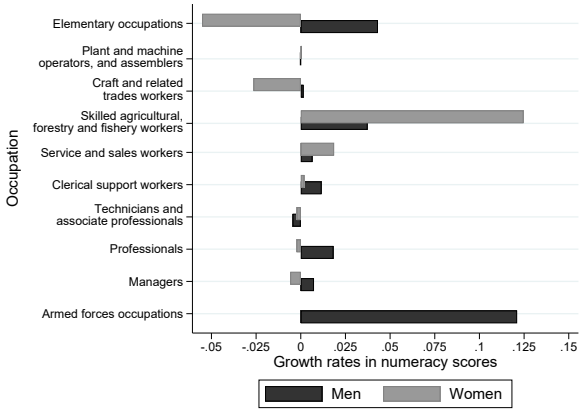
(C) By education level 2012

Notes: Growth rates in numeracy scores for men and women in Germany between 2015 and 2012 by numeracy quartiles in 2012 (A), number of children (B), and education level (C), all in 2012. Growth rates are calculated by dividing the difference between 2015 and 2012 numeracy scores by 2012 numeracy scores. Age groups refer to the age reported in 2012. Confidence intervals are added for each bar. Sample contains all individuals with non-missing numeracy scores in 2012 and 2015, as well as the respective categories (A: 2,997 individuals, B: 2,961 individuals, C: 2,960 individuals). Data source: PIAAC-L German SUF 2015 and 2012.

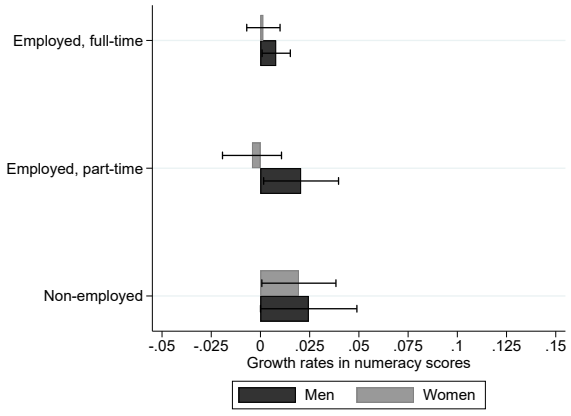
Figure A5.10 : Change in Numeracy Test Score between 2012 and 2015 (II)



(A) By field of study 2012



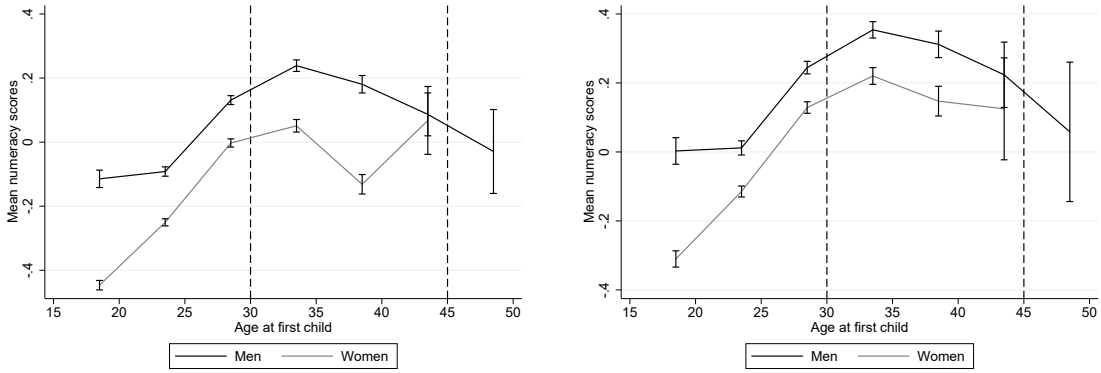
(B) By occupational category 2012



(C) By employment status 2012

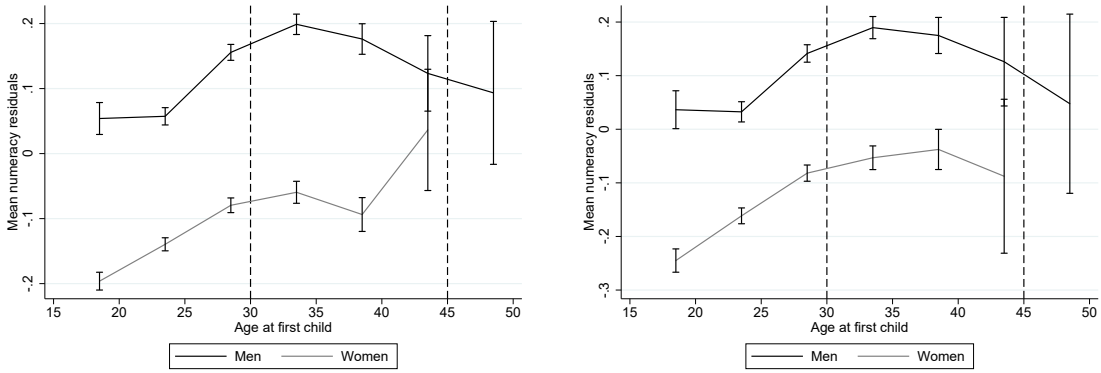
Notes: Growth rates in numeracy scores for men and women in Germany between 2015 and 2012 by field of study (A), number of occupation (B), and employment status (C), all in 2012. Growth rates are calculated by dividing the difference between 2015 and 2012 numeracy scores by 2012 numeracy scores. Age groups refer to the age reported in 2012. Confidence intervals are added for each bar. Sample contains all individuals with non-missing numeracy scores in 2012 and 2015, as well as the respective categories (A: 2,959 individuals, B: 2,355 individuals, C: 2,950 individuals). Data source: PIAAC-L German SUF 2015 and 2012.

Figure A5.11 : Numeracy Levels by Gender and Age at the first Childbirth



(a) All individuals (a) Non-missing wages

(A) Unconditional

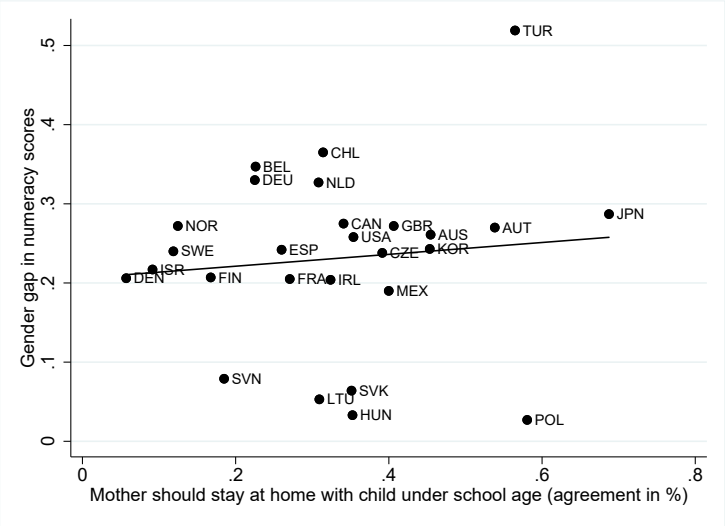


(a) All individuals (a) Non-missing wages

(B) Residual, after conditioning on educational levels

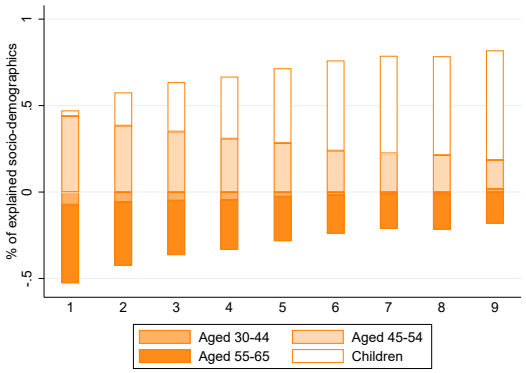
Notes: Mean standardized numeracy scores by age at birth of first child (in five-year intervals) for men and women aged 20 to 65. Panel (A) presents raw numeracy scores, panel (B) plots the residuals of a least squares regression of numeracy scores on age groups, education levels, and country dummies, using sampling weights. Confidence intervals for each data point are added, vertical lines represent cut-offs of age groups used in the regressions at ages 30, 45, and 55. Standardization by country uses individuals' sampling probability. Sample contains all individuals with non-missing numeracy scores, age, and child information (unconditional A: 139,104 individuals; B: 67,243 individuals, residual A: 139,018 individuals; B: 67,202 individuals). Data source: PIAAC international PUF 2012.

Figure A5.12 : Relationship between Numeracy Gaps and Norms, by Country

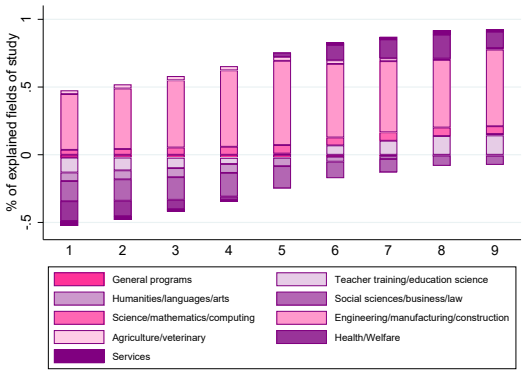


Notes: Gender pay gap in standardized numeracy scores for men and women aged 20 to 65 by country plotted against the percentage in agreement to the statement "Do you think that women should work outside the home full-time, part-time or not at all under the following circumstances?" Option: "Stay at home when there is a child under school age." by country, including a linear fit. The R-squared from a simple regression of gender numeracy gaps on gender norms is 0.27. Sample contains individuals aged 20 to 65 with non-missing numeracy scores from PIAAC and countries with non-missing norms information from ISSP 2012 (27 countries, i.e. 165,961 individuals). Data source: PIAAC international PUF 2012 and data on norms from the 2012 ISSP questionnaire on "Family and Changing Gender Roles" (ISSP Research Group, 2016).

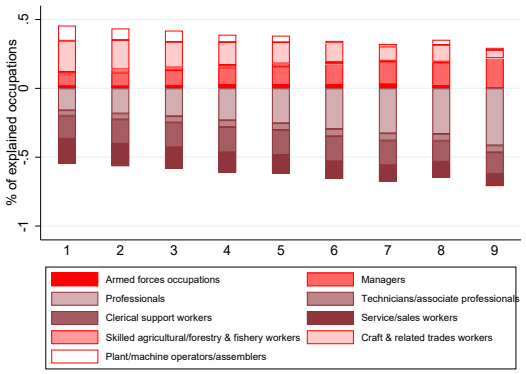
Figure A5.13 : Decomposition of the Gender Numeracy Gap: Explained Part, Selected Groups



(A) Explained: Socio-demographics



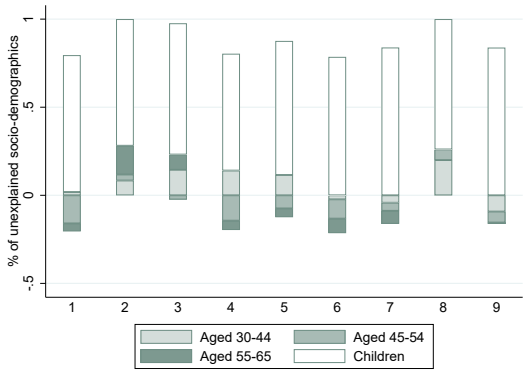
(B) Explained: Field of study



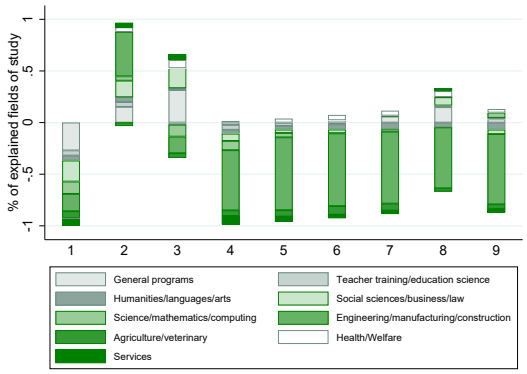
(C) Explained: Occupation

Notes: Explained part of a detailed Kitagawa–Oaxaca–Blinder type decomposition of gender numeracy gaps by numeracy decile using the command *oaxaca_rif*. Explanatory variables presented here: age groups and children (A), field of study (B), and occupation (C). Results look similar when just considering countries with earnings information or only individuals with non-missing wages (results available upon request). Sample contains all individuals with non-missing numeracy and the respective controls (144,371 individuals). Data source: PIAAC international PUF 2012.

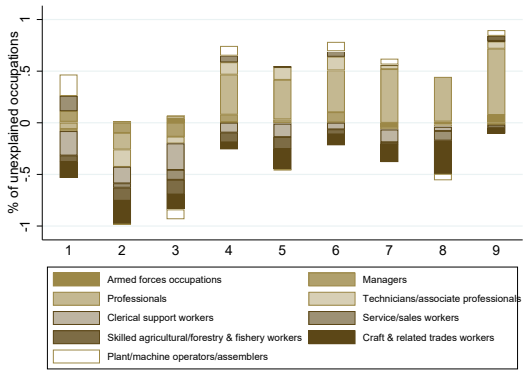
Figure A5.14 : Decomposition of the Gender Numeracy Gap: Unexplained Part, Selected Groups



(A) Unexplained: Socio-demographics



(B) Unexplained: Field of study



(C) Unexplained: Occupation

Notes: Unexplained part of a detailed Kitagawa–Oaxaca–Blinder type decomposition of gender numeracy gaps by numeracy decile using the command *oaxaca_rif*. Explanatory variables presented here: age groups and children (A), field of study (B), and occupation (C). Results differ slightly when just considering countries with earnings information or only individuals with non-missing wages (results available upon request). Sample contains all individuals with non-missing numeracy and the respective controls (144,371 individuals). Data source: PIAAC international PUF 2012.

Table A5.1 : Composition of PIAAC Data by Country

Country	Isocode	2011/12	2014/15	2017	Numeracy	Literacy	Problem solving	Wages
Australia	AUS	x			6,974	6,974	5,217	4,266
Austria	AUT	x			4,597	4,597	3,451	2,824 (D)
Belgium	BEL	x			4,542	4,542	3,755	2,751
Canada	CAN	x			24,462	24,462	19,183	15,248 (D)
Chile	CHL		x		4,770	4,770	2,954	2,298
Cyprus	CYP	x			4,093	4,093	0	2,149
Czech Republic	CZE	x			5,357	5,357	3,984	2,581
Denmark	DEN	x			6,770	6,770	5,620	4,447
Ecuador	ECU			x	4,964	4,964	1,991	1,652
Estonia	EST	x			7,043	7,043	4,715	3,999
Finland	FIN	x			5,042	5,042	4,100	3,252
France	FRA	x			6,374	6,374	0	3,719
Germany	DEU	x			4,871	4,871	4,049	3,279
Greece	GRC		x		4,684	4,684	2,965	1,260
Hungary	HUN			x	5,719	5,719	3,700	0
Ireland	IRL	x			5,626	5,626	3,788	2,788
Israel	ISR		x		4,722	4,722	3,123	2,605
Italy	ITA	x			4,367	4,367	0	1,978
Japan	JPN	x			4,806	4,806	3,034	3,239
Kazakhstan	KAZ			x	5,706	5,706	4,205	2,680
Korea	KOR	x			6,081	6,081	3,998	3,095
Lithuania	LTU		x		4,783	4,783	3,421	2,746
Mexico	MEX			x	5,616	5,616	2,008	2,253
Netherlands	NLD	x			4,655	4,655	4,139	2,997
New Zealand	NZL		x		5,457	5,457	4,922	3,314
Norway	NOR	x			4,455	4,455	3,872	3,408
Peru	PER			x	6,538	6,538	2,867	0
Poland	POL	x			8,302	8,302	5,129	3,839
Singapore	SGP		x		4,887	4,887	3,598	0
Slovak Republic	SVK	x			5,213	5,213	3,110	2,510
Slovenia	SVN		x		4,922	4,922	3,633	2,233
Spain	ESP	x			5,504	5,504	0	2,471
Sweden	SWE	x			4,080	4,080	3,591	2,872 (D)
Turkey	TUR		x		4,854	4,854	2,038	0
United Kingdom	GBR	x			8,311	8,311	6,850	4,728
United States	USA	x		x	4,553	4,553	3,786	2,734 (D)
Total	36	23	8	6	213,700	213,700	136,796	106,215

Notes: The table contains the list of participating countries and their ISO codes, and an indication of the year when the survey was conducted (the first round in 2011/12, the second round in 2014/15, or the third round in 2017). Additionally, the table lists the number of non-missing observations available for each of the skill domains (numeracy, literacy, problem solving) and wages. (D) denotes countries that provide wage information only by belonging to a decile. Also note that the list does not include Russia and Indonesia, following the recommendation in the official PIAAC reports. For details also see Appendix A5.1 and `https://www.oecd.org/skills/piaac/about`. Data source: PIAAC international PUF 2012, own calculations.

Table A5.2 : Returns to Numeracy Levels (no Further Controls)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.168***	-0.147***	-0.168***	-0.152***	-0.154***	-0.136***	-0.121***	-0.098***	-0.093***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Aged 30-44	0.147***	0.177***	0.231***	0.266***	0.300***	0.300***	0.296***	0.279***	0.248***
	(0.014)	(0.010)	(0.009)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.008)
Aged 45-54	0.098***	0.146***	0.219***	0.283***	0.341***	0.367***	0.381***	0.375***	0.366***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Aged 55-65	0.067***	0.138***	0.209***	0.271***	0.337***	0.368***	0.387***	0.390***	0.400***
	(0.017)	(0.012)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.012)
Numeracy above median	0.191***	0.224***	0.232***	0.232***	0.254***	0.274***	0.297***	0.334***	0.365***
	(0.014)	(0.010)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.010)
Numeracy above median × Female	0.073***	0.035*	0.036**	0.024*	0.020*	-0.002	-0.028**	-0.082***	-0.139***
	(0.020)	(0.014)	(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.013)
Education levels	No	No	No	No	No	No	No	No	No
Field of study	No	No	No	No	No	No	No	No	No
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	106206	106206	106206	106206	106206	106206	106206	106206	106206

Notes: Corresponding table for figure 5.4. Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.3 : Returns to Numeracy Levels (controlling for Education and Field of Study)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.221***	-0.191***	-0.204***	-0.193***	-0.198***	-0.180***	-0.162***	-0.141***	-0.129***
	(0.017)	(0.012)	(0.010)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)	(0.009)
Aged 30-44	0.119***	0.144***	0.196***	0.240***	0.270***	0.271***	0.269***	0.251***	0.214***
	(0.014)	(0.010)	(0.009)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.008)
Aged 45-54	0.095***	0.137***	0.207***	0.279***	0.336***	0.365***	0.383***	0.376***	0.363***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Aged 55-65	0.067***	0.130***	0.198***	0.270***	0.333***	0.367***	0.392***	0.394***	0.404***
	(0.017)	(0.012)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.012)
Numeracy above median	0.074***	0.118***	0.126***	0.126***	0.137***	0.154***	0.176***	0.203***	0.220***
	(0.014)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.008)	(0.010)
Numeracy above median × Female	0.082***	0.035*	0.032**	0.026*	0.028**	0.005	-0.024*	-0.077***	-0.130***
	(0.020)	(0.014)	(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.013)
Education levels	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field of study	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	103443	103443	103443	103443	103443	103443	103443	103443	103443

Notes: Corresponding table for figure A5.8 (A). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.4 : Returns to Numeracy Levels (controlling for Education, Field of Study, and Occupation)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.161***	-0.168***	-0.181***	-0.165***	-0.168***	-0.151***	-0.141***	-0.125***	-0.124***
	(0.017)	(0.012)	(0.010)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.010)
Aged 30-44	0.087***	0.116***	0.167***	0.209***	0.237***	0.237***	0.235***	0.216***	0.179***
	(0.014)	(0.010)	(0.009)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.008)
Aged 45-54	0.058***	0.104***	0.172***	0.242***	0.296***	0.323***	0.340***	0.331***	0.318***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Aged 55-65	0.030	0.095***	0.159***	0.229***	0.289***	0.320***	0.345***	0.343***	0.349***
	(0.017)	(0.012)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.011)
Numeracy above median	0.038*	0.072***	0.081***	0.081***	0.088***	0.101***	0.119***	0.145***	0.160***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.008)	(0.010)
Numeracy above median × Female	0.046*	0.027	0.023	0.014	0.017	-0.000	-0.021*	-0.067***	-0.111***
	(0.020)	(0.015)	(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.013)
Education levels	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field of study	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Full-time indicator	No	No	No	No	No	No	No	No	No
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	102602	102602	102602	102602	102602	102602	102602	102602	102602

Notes: Corresponding table for figure A5.8 (B). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.5 : Returns to Numeracy Levels (controlling for Education, Field of Study, Occupation, and Full-Time Status)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.200***	-0.200***	-0.200***	-0.173***	-0.173***	-0.155***	-0.149***	-0.143***	-0.158***
	(0.018)	(0.012)	(0.010)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.010)
Aged 30-44	0.102***	0.128***	0.174***	0.213***	0.239***	0.239***	0.238***	0.222***	0.191***
	(0.014)	(0.010)	(0.009)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.008)
Aged 45-54	0.072***	0.116***	0.179***	0.246***	0.298***	0.325***	0.343***	0.337***	0.330***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Aged 55-65	0.030	0.096***	0.160***	0.230***	0.289***	0.320***	0.345***	0.343***	0.350***
	(0.017)	(0.012)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.011)
Numeracy above median	0.035*	0.069***	0.080***	0.081***	0.087***	0.101***	0.118***	0.143***	0.157***
	(0.015)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.008)	(0.010)
Numeracy above median × Female	0.054**	0.034*	0.027*	0.016	0.018	0.001	-0.019*	-0.064***	-0.104***
	(0.020)	(0.014)	(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.013)
Education levels	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field of study	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Full-time indicator	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	102567	102567	102567	102567	102567	102567	102567	102567	102567

Notes: Corresponding table for figure A5.8 (C). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.6 : Returns to Numeracy Levels (controlling for Education, Field of Study, Occupation, Full-time Status, and Children)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.135***	-0.127***	-0.146***	-0.127***	-0.125***	-0.108***	-0.094***	-0.066***	-0.051***
	(0.023)	(0.018)	(0.014)	(0.012)	(0.011)	(0.011)	(0.010)	(0.010)	(0.013)
Aged 30-44	0.099***	0.116***	0.160***	0.194***	0.213***	0.211***	0.209***	0.192***	0.157***
	(0.015)	(0.011)	(0.010)	(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.009)
Aged 45-54	0.068***	0.099***	0.160***	0.221***	0.264***	0.289***	0.306***	0.298***	0.285***
	(0.018)	(0.013)	(0.011)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)	(0.012)
Aged 55-65	0.025	0.077***	0.140***	0.203***	0.252***	0.282***	0.305***	0.300***	0.300***
	(0.020)	(0.014)	(0.011)	(0.011)	(0.009)	(0.009)	(0.009)	(0.010)	(0.014)
Numeracy above median	0.035*	0.071***	0.081***	0.082***	0.088***	0.102***	0.119***	0.145***	0.159***
	(0.014)	(0.011)	(0.010)	(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.010)
Numeracy above median × Female	0.050*	0.029	0.024*	0.013	0.016	-0.001	-0.022*	-0.068***	-0.111***
	(0.020)	(0.015)	(0.012)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.012)
Children	0.050**	0.076***	0.067***	0.072***	0.088***	0.091***	0.099***	0.116***	0.146***
	(0.016)	(0.011)	(0.010)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.011)
Children × Female	-0.094***	-0.107***	-0.081***	-0.070***	-0.075***	-0.075***	-0.086***	-0.116***	-0.161***
	(0.021)	(0.015)	(0.014)	(0.011)	(0.011)	(0.010)	(0.010)	(0.011)	(0.014)
Education levels	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field of study	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Full-time indicator	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	102506	102506	102506	102506	102506	102506	102506	102506	102506

Notes: Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.7 : Returns to Numeracy Levels for those Without Children (no Further Controls)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.061*	-0.034	-0.062***	-0.046***	-0.051***	-0.058***	-0.052***	-0.039**	-0.020
	(0.026)	(0.018)	(0.016)	(0.013)	(0.013)	(0.012)	(0.012)	(0.013)	(0.016)
Aged 30-44	0.160***	0.173***	0.216***	0.233***	0.277***	0.281***	0.307***	0.320***	0.286***
	(0.016)	(0.013)	(0.011)	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)	(0.014)
Aged 45-54	0.098***	0.110***	0.186***	0.231***	0.309***	0.353***	0.384***	0.452***	0.476***
	(0.024)	(0.018)	(0.015)	(0.012)	(0.013)	(0.013)	(0.014)	(0.017)	(0.023)
Aged 55-65	0.046	0.118***	0.190***	0.243***	0.292***	0.341***	0.383***	0.439***	0.519***
	(0.032)	(0.022)	(0.019)	(0.016)	(0.017)	(0.018)	(0.020)	(0.024)	(0.034)
Numeracy above median	0.182***	0.220***	0.203***	0.201***	0.214***	0.209***	0.227***	0.256***	0.257***
	(0.022)	(0.016)	(0.014)	(0.011)	(0.011)	(0.011)	(0.011)	(0.013)	(0.016)
Numeracy above median × Female	0.040	-0.010	0.014	-0.001	-0.000	0.012	0.006	-0.023	-0.049*
	(0.031)	(0.023)	(0.020)	(0.016)	(0.017)	(0.016)	(0.016)	(0.018)	(0.023)
Education levels	No	No	No	No	No	No	No	No	No
Field of study	No	No	No	No	No	No	No	No	No
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	34645	34645	34645	34645	34645	34645	34645	34645	34645

Notes: Corresponding table for figure 5.7 (A). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.8 : Returns to Numeracy Levels for those With Children (no Further Controls)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.196***	-0.208***	-0.222***	-0.229***	-0.198***	-0.167***	-0.136***	-0.116***	-0.097***
	(0.019)	(0.013)	(0.011)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)
Aged 30-44	0.151***	0.154***	0.177***	0.220***	0.225***	0.225***	0.222***	0.205***	0.179***
	(0.029)	(0.020)	(0.017)	(0.015)	(0.013)	(0.011)	(0.010)	(0.010)	(0.010)
Aged 45-54	0.100***	0.124***	0.169***	0.240***	0.268***	0.291***	0.301***	0.284***	0.257***
	(0.030)	(0.020)	(0.018)	(0.015)	(0.014)	(0.012)	(0.011)	(0.011)	(0.011)
Aged 55-65	0.078*	0.115***	0.151***	0.225***	0.263***	0.288***	0.300***	0.294***	0.298***
	(0.031)	(0.021)	(0.018)	(0.016)	(0.014)	(0.013)	(0.012)	(0.012)	(0.013)
Numeracy above median	0.201***	0.223***	0.252***	0.266***	0.293***	0.314***	0.350***	0.377***	0.374***
	(0.017)	(0.013)	(0.011)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.012)
Numeracy above median × Female	0.069**	0.058**	0.038*	0.036**	0.006	-0.030**	-0.071***	-0.130***	-0.178***
	(0.024)	(0.018)	(0.015)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)
Education levels	No	No	No	No	No	No	No	No	No
Field of study	No	No	No	No	No	No	No	No	No
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71489	71489	71489	71489	71489	71489	71489	71489	71489

Notes: Corresponding table for figure 5.7 (B). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.9 : Returns to Numeracy Levels, non-STEM Field of Study (no Further Controls)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.109***	-0.100***	-0.130***	-0.107***	-0.115***	-0.094***	-0.095***	-0.079***	-0.091***
	(0.019)	(0.013)	(0.011)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.011)
Aged 30-44	0.132***	0.163***	0.205***	0.252***	0.290***	0.294***	0.290***	0.276***	0.241***
	(0.017)	(0.012)	(0.010)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.009)
Aged 45-54	0.077***	0.145***	0.199***	0.264***	0.325***	0.358***	0.375***	0.368***	0.361***
	(0.019)	(0.013)	(0.011)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.011)
Aged 55-65	0.061**	0.132***	0.188***	0.257***	0.329***	0.367***	0.390***	0.389***	0.397***
	(0.021)	(0.014)	(0.012)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.014)
Numeracy above median	0.192***	0.224***	0.219***	0.229***	0.252***	0.275***	0.295***	0.332***	0.365***
	(0.019)	(0.013)	(0.011)	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.014)
Numeracy above median × Female	0.046*	0.028	0.022	0.006	0.009	-0.008	-0.026*	-0.082***	-0.136***
	(0.025)	(0.017)	(0.014)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.017)
Education levels	No	No	No	No	No	No	No	No	No
Field of study	No	No	No	No	No	No	No	No	No
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	76961	76961	76961	76961	76961	76961	76961	76961	76961

Notes: Corresponding table for figure 5.8 (A). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Table A5.10 : Returns to Numeracy Levels, STEM Field of Study (no Further Controls)

	Outcome: Log Hourly Wages								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	-0.409***	-0.351***	-0.238***	-0.215***	-0.193***	-0.133***	-0.094***	-0.072***	-0.044**
	(0.050)	(0.034)	(0.023)	(0.021)	(0.018)	(0.016)	(0.016)	(0.016)	(0.015)
Aged 30-44	0.153***	0.195***	0.243***	0.294***	0.291***	0.296***	0.295***	0.278***	0.203***
	(0.024)	(0.020)	(0.017)	(0.015)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)
Aged 45-54	0.093***	0.142***	0.240***	0.344***	0.363***	0.383***	0.387***	0.381***	0.325***
	(0.026)	(0.023)	(0.019)	(0.016)	(0.014)	(0.013)	(0.013)	(0.014)	(0.015)
Aged 55-65	0.040	0.140***	0.191***	0.294***	0.315***	0.348***	0.377***	0.379***	0.348***
	(0.032)	(0.026)	(0.021)	(0.018)	(0.016)	(0.016)	(0.016)	(0.017)	(0.019)
Numeracy above median	0.202***	0.222***	0.216***	0.226***	0.220***	0.245***	0.288***	0.302***	0.246***
	(0.021)	(0.018)	(0.014)	(0.013)	(0.011)	(0.011)	(0.011)	(0.012)	(0.013)
Numeracy above median × Female	0.274***	0.202***	0.102***	0.068**	0.050*	-0.008	-0.051*	-0.068**	-0.097***
	(0.056)	(0.040)	(0.029)	(0.026)	(0.023)	(0.021)	(0.021)	(0.022)	(0.022)
Education levels	No	No	No	No	No	No	No	No	No
Field of study	No	No	No	No	No	No	No	No	No
Occupation	No	No	No	No	No	No	No	No	No
Full-time indicator	No	No	No	No	No	No	No	No	No
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	26496	26496	26496	26496	26496	26496	26496	26496	26496

Notes: Corresponding table for figure 5.8 (B). Dependent Variable: (log) trimmed gross hourly wages (ppp-adjusted). Wage measures are trimmed and imputed with decile medians if the continuous measure was not available. Numeracy skill measures are standardized at the country level using sampling probabilities. Unconditional quantile regression with controls for education, field of study, occupation, a full-time indicator and country fixed effects at each wage decile, weighted by individual sampling probability. Estimation sample contains all individuals with non-missing data for wages and respective controls. Data source: PIAAC international PUF 2012.

Bibliography

Abramitzky, Ran and Leah Boustan (2017). "Immigration in American economic history". *Journal of Economic Literature* 55(4), 1311–1345.

Acemoglu, Daron and David Autor (2011). "Skills, tasks and technologies: Implications for employment and earnings". *Handbook of Labor Economics*. Vol. 4. Elsevier, 1043–1171.

Adda, Jérôme, Christian Dustmann, and Katrien Stevens (2017). "The career costs of children". *Journal of Political Economy* 125(2), 293–337.

Aguiar, Luis, Christian Peukert, Maximilian Schäfer, and Hannes Ullrich (2022). "Facebook shadow profiles". *arXiv preprint arXiv:2202.04131*.

Akyol, Pelin, Kala Krishna, and Jinwen Wang (2021). "Taking PISA seriously: How accurate are low-stakes exams?" *Journal of Labor Research* 42, 184–243.

Alan, Sule, Nazli Baydar, Teodora Boneva, Thomas F Crossley, and Seda Ertac (2017). "Transmission of risk preferences from mothers to daughters". *Journal of Economic Behavior & Organization* 134, 60–77.

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

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

Albrecht, James, Anders Bjorklund, and Susan Vroman (2003). "Is There a Glass Ceiling in Sweden?" *Journal of Labor Economics* 21(1), 145–177.

Alesina, Alberto and Paola Giuliano (2014). "Family ties". *Handbook of Economic Growth*. Vol. 2. Elsevier, 177–215.

– (2015). "Culture and institutions". *Journal of Economic Literature* 53(4), 898–944.

Allgood, Sam, Lee Badgett, Amanda Bayer, Marianne Bertrand, Sandra E. Black, Nick Bloom, and Lisa D. Cook (2019). "AEA Professional Climate Survey: Final Report". *American Economic Assocation*.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz (2011). "Personality psychology and economics". *Handbook of the Economics of Education*. Vol. 4. Elsevier, 1–181.

Ammons, Robert B (1956). "Effects of knowledge of performance: A survey and tentative theoretical formulation". *The Journal of General Psychology* 54(2), 279–299.

Andreoni, James and Charles Sprenger (2012). "Risk preferences are not time preferences". *American Economic Review* 102(7), 3357–3376.

Andreoni, James and Charles Sprenger (2015). “Risk preferences are not time preferences: Reply”. *American Economic Review* 105(7), 2287–2293.

Angerer, Silvia, Jana Bolvashenkova, Daniela Glätzle-Rützle, Philipp Lergetporer, and Matthias Sutter (2023). “Children’s patience and school-track choices several years later: Linking experimental and field data”. *Journal of Public Economics* forthcoming.

Arellano-Bover, Jaime (2022). “The effect of labor market conditions at entry on workers’ long-term skills”. *Review of Economics and Statistics* 104(5), 1028–1045.

Arulampalam, Wiji, Alison L. Booth, and Mark L. Bryan (2007). “Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wage Distribution”. *ILR Review* 60(2), 163–186.

Autor, David, David N. Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman (2020). *Males at the Tails: How Socioeconomic Status Shapes the Gender Gap*. NBER Working Papers 27196. National Bureau of Economic Research, Inc.

Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri (2019). “What you don’t know... can’t hurt you? A natural field experiment on relative performance feedback in higher education”. *Management Science* 65(8), 3714–3736.

Azmat, Ghazala and Nagore Iriberri (2010). “The importance of relative performance feedback information: Evidence from a natural experiment using high school students”. *Journal of Public Economics* 94(7-8), 435–452.

Bacolod, Marigee P. and Bernardo S. Blum (2010). “Two Sides of the Same Coin: U.S. "Residual" Inequality and the Gender Gap”. *The Journal of Human Resources* 45(1), 197–242.

Bailey, Michael, Drew M Johnston, Martin Koenen, Theresa Kuchler, Dominic Russel, and Johannes Stroebel (2022). “The social integration of international migrants: Evidence from the networks of Syrians in Germany”. *NBER Working Paper 29925*.

Balart, Pau, Matthijs Oosterveen, and Dinand Webbink (2018). “Test scores, noncognitive skills and economic growth”. *Economics of Education Review* 63, 134–153.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul (2015). “Blissful ignorance? A natural experiment on the effect of feedback on students’ performance”. *Labour Economics* 34, 13–25.

Baumeister, Roy F, Kathleen D Vohs, C Nathan DeWall, and Liqing Zhang (2007). “How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation”. *Personality and Social Psychology Review* 11(2), 167–203.

Becker, Gary S (1962). “Investment in human capital: A theoretical analysis”. *Journal of Political Economy* 70(5, Part 2), 9–49.

– (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. New York, NY: National Bureau of Economic Research.

Ben-Porath, Yoram (1967). “The production of human capital and the life cycle of earnings”. *Journal of Political Economy* 75(4, Part 1), 352–365.

Ben-Porath, Yorman (1970). “The production of human capital over time”. *Education, income, and human capital*. NBER, 129–154.

Bénabou, Roland and Jean Tirole (2002). “Self-confidence and personal motivation”. *The Quarterly Journal of Economics* 117(3), 871–915.

Benzoni, Luca and Olena Chyruk (2015). “The value and risk of human capital”. *Annual Review of Financial Economics* 7, 179–200.

Bertrand, Marianne (2018). “Coase Lecture – The Glass Ceiling”. *Economica* 85(338), 205–231.

– (2020). “Gender in the 21st century”. AEA Distinguished Lecture Series.

Bird, Edward J (2001). “Does the welfare state induce risk-taking?” *Journal of Public Economics* 80(3), 357–383.

Bisin, Alberto and Thierry Verdier (2011). “The economics of cultural transmission and socialization”. *Handbook of Social Economics*. Vol. 1. Elsevier, 339–416.

Blau, Francine D. and Lawrence M. Kahn (2017). “The Gender Wage Gap: Extent, Trends, and Explanations”. *Journal of Economic Literature* 55(3), 789–865.

Borghans, Lex and Trudie Schils (2012). “The leaning tower of Pisa: Decomposing achievement test scores into cognitive and noncognitive components”. *Unpublished manuscript*.

Borjas, George J (1987). *Self-selection and the earnings of immigrants*. Tech. rep. 4, 531–553.

Brade, Raphael, Oliver Himmler, and Robert Jäckle (2022). “Relative performance feedback and the effects of being above average - Field experiment and replication”. *Economics of Education Review* 89, 102268.

Brown, Jeffrey, Chichun Fang, and Francisco Gomes (2012). “Risk and returns to education”. *NBER Working Paper 18300*.

Bursztyn, Leonardo and Robert Jensen (2015). “How does peer pressure affect educational investments?” *The Quarterly Journal of Economics* 130(3), 1329–1367.

Buser, Thomas (2016). “The impact of losing in a competition on the willingness to seek further challenges”. *Management Science* 62(12), 3439–3449.

Cabañas, Jose González, Ángel Cuevas, and Rubén Cuevas (2018). “Facebook use of sensitive data for advertising in Europe”. *arXiv preprint arXiv:1802 .05030*.

Cabus, Sofie J and Kristof De Witte (2016). “Why do students leave education early? Theory and evidence on high school dropout rates”. *Journal of Forecasting* 35(8), 690–702.

Carroll, Christopher D, Byung-Kun Rhee, and Changyong Rhee (1994). “Are there cultural effects on saving? Some cross-sectional evidence”. *The Quarterly Journal of Economics* 109(3), 685–699.

Castillo, Marco, Jeffrey L Jordan, and Ragan Petrie (2018). “Children’s rationality, risk attitudes and field behavior”. *European Economic Review* 102, 62–81.

– (2019). “Discount rates of children and high school graduation”. *The Economic Journal* 129(619), 1153–1181.

Castillo, Marco, John A List, Ragan Petrie, and Anya Samek (2020). "Detecting drivers of behavior at an early age: Evidence from a longitudinal field experiment". *NBER Working Paper 28288*.

Chetty, Raj, Matthew O Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, et al. (2022). "Social capital I: Measurement and associations with economic mobility". *Nature* 608(7921), 108–121.

Cheung, Stephen L (2015). "Comment on "Risk preferences are not time preferences": On the elicitation of time preference under conditions of risk". *American Economic Review* 105(7), 2242–2260.

Choi, Eunju, Douglas A Johnson, Kwangsu Moon, and Shezeen Oah (2018). "Effects of positive and negative feedback sequence on work performance and emotional responses". *Journal of Organizational Behavior Management* 38(2-3), 97–115.

Christl, M. and Monika Köppl-Turyna (2020). "Gender wage gap and the role of skills and tasks: Evidence from the Austrian PIAAC data set". *Applied Economics* 52(2), 113–134.

Coffman, Katherine B, Paola Ugalde Araya, and Basit Zafar (2021). "A (Dynamic) Investigation of Stereotypes, Belief-Updating, and Behavior". *NBER Working Paper 29382*.

Collischon, Matthias (2019). "Is There a Glass Ceiling over Germany?" *German Economic Review* 20(4), 329–359.

Contini, Dalit, Maria Laura Di Tommaso, and Silvia Mendolia (2017). "The gender gap in mathematics achievement: Evidence from Italian data". *Economics of Education Review* 58(C), 32–42.

Cordero, Jose M, Cristina Polo, Daniel Santıén, and Rosa Simancas (2018). "Efficiency measurement and cross-country differences among schools: A robust conditional nonparametric analysis". *Economic Modelling* 74, 45–60.

Cunha, Flavio, James J Heckman, and Susanne M Schennach (2010). "Estimating the technology of cognitive and noncognitive skill formation". *Econometrica* 78(3), 883–931.

Davies, Don and Alfred Jacobs (1985). "'Sandwiching' complex interpersonal feedback". *Small Group Behavior* 16(3), 387–396.

De Philippis, Marta and Federico Rossi (2021). "Parents, schools and human capital differences across countries". *Journal of the European Economic Association* 19(2), 1364–1406.

Dee, Thomas S and Brian Jacob (2011). "The impact of No Child Left Behind on student achievement". *Journal of Policy Analysis and Management* 30(3), 418–446.

Dinerstein, Michael, Rigissa Megalokonomou, Constantine Yannelis, et al. (2022). "Human Capital Depreciation and Returns to Experience". *American Economic Review* 112(11), 3725–3762.

Dobrescu, LI, Marco Faravelli, Rigissa Megalokonomou, and Alberto Motta (2021). "Relative performance feedback in education: Evidence from a randomised controlled trial". *The Economic Journal* 131(640), 3145–3181.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde (2010). "Are risk aversion and impatience related to cognitive ability?" *American Economic Review* 100(3), 1238–1260.

– (2018). "On the relationship between cognitive ability and risk preference". *Journal of Economic Perspectives* 32(2), 115–34.

Duckworth, Angela L, Christopher Peterson, Michael D Matthews, and Dennis R Kelly (2007). "Grit: perseverance and passion for long-term goals." *Journal of Personality and Social Psychology* 92(6), 1087.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, et al. (2021). "Gender and the dynamics of Economics seminars". *NBER Working Paper 28494*.

Edin, Per-Anders and Magnus Gustavsson (2008). "Time Out of Work and Skill Depreciation". *ILR Review* 61(2), 163–180.

Eil, David and Justin M Rao (2011). "The good news-bad news effect: Asymmetric processing of objective information about yourself". *American Economic Journal: Microeconomics* 3(2), 114–38.

Ellison, Glenn and Ashley Swanson (2010). "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions". *Journal of Economic Perspectives* 24(2), 109–128.

Epper, Thomas and Helga Fehr-Duda (2015). "Comment on "Risk preferences are not time preferences": Balancing on a budget line". *American Economic Review* 105(7), 2261–2271.

Erickson, Devon, D Kip Holderness Jr, Kari Joseph Olsen, and Todd A Thornock (2021). "Feedback with feeling? How emotional language in feedback affects individual performance". *Accounting, Organizations and Society*, 101329.

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

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (2018). "Global evidence on economic preferences". *The Quarterly Journal of Economics* 133(4), 1645–1692.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde (2022). "The preference survey module: A validated instrument for measuring risk, time, and social preferences". *Management Science, forthcoming*.

Falorsi, Piero Demetrio, Roberto Ricci, and Patrizia Falzetti (2019). *Le metodologie di campionamento e scomposizione della devianza nelle rilevazioni nazionali dell'INVALSI: Le rilevazioni degli apprendimenti AS 2018-2019*. Franco Angeli.

Fernández, Raquel and Alessandra Fogli (2009). "Culture: An empirical investigation of beliefs, work, and fertility". *American Economic Journal: Macroeconomics* 1(1), 146–177.

Figlio, David, Paola Giuliano, Umut Özek, and Paola Sapienza (2019). "Long-term orientation and educational performance". *American Economic Journal: Economic Policy* 11(4), 272–309.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009). "Unconditional Quantile Regressions". *Econometrica* 77(3), 953–973.

Fischer, Mira and Valentin Wagner (2018). "Effects of timing and reference frame of feedback: Evidence from a field experiment". *IZA Discussion Paper 11970*.

Freeman, Richard B (1999). "The Economics of crime". *Handbook of Labor Economics* 3, 3529–3571.

Galor, Oded and Ömer Özak (2016). "The agricultural origins of time preference". *American Economic Review* 106(10), 3064–3103.

Gillet, Nicolas, Sophie Berjot, Robert J Vallerand, and Sofiane Amoura (2012). "The role of autonomy support and motivation in the prediction of interest and dropout intentions in sport and education settings". *Basic and Applied Social Psychology* 34(3), 278–286.

Giuliano, Paola (2007). "Living arrangements in Western Europe: Does cultural origin matter?" *Journal of the European Economic Association* 5(5), 927–952.

Gneezy, Uri, John A List, Jeffrey A Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu (2019). "Measuring success in education: The role of effort on the test itself". *American Economic Review: Insights* 1(3), 291–308.

Goldin, Claudia (2014). "A grand gender convergence: Its last chapter". *American Economic Review* 104(4), 1091–1119.

Golsteyn, Bart HH, Hans Grönqvist, and Lena Lindahl (2014). "Adolescent time preferences predict lifetime outcomes". *The Economic Journal* 124(580), F739–F761.

Goulas, Sofoklis, Silvia Griselda, and Rigissa Megalokonomou (2022). "Comparative advantage and gender gap in STEM". *Journal of Human Resources*, 0320–10781R2.

Goulas, Sofoklis and Rigissa Megalokonomou (2021). "Knowing who you actually are: The effect of feedback on short-and longer-term outcomes". *Journal of Economic Behavior & Organization* 183, 589–615.

Griselda, Silvia (2022). "The Gender Gap in Math: What are We Measuring?" *Available at SSRN 4022082*.

Grogger, Jeffrey and Gordon H Hanson (2011). "Income maximization and the selection and sorting of international migrants". *Journal of Development Economics* 95(1), 42–57.

Groot, Wim and Hessel Oosterbeek (1992). "Optimal investment in human capital under uncertainty". *Economics of Education Review* 11(1), 41–49.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2004). "The role of social capital in financial development". *American Economic Review* 94(3), 526–556.

– (2006). "Does culture affect economic outcomes?" *Journal of Economic Perspectives* 20(2), 23–48.

Halevy, Yoram (2008). "Strotz meets Allais: Diminishing impatience and the certainty effect". *American Economic Review* 98(3), 1145–1162.

Halpern, Diane F (2013). *Sex differences in cognitive abilities*. Psychology press.

Hampf, Franziska, Marc Piopiunik, and Simon Wiederhold (2020). *The Effects of Graduating from High School in a Recession: College Investments, Skill Formation, and Labor-Market Outcomes*. CESifo Working Paper Series 8252. CESifo.

Hanushek, Eric A (1986). “The economics of schooling: Production and efficiency in public schools”. *Journal of Economic Literature* 24(3), 1141–1177.

– (2016). “What matters for achievement: updating Coleman on the influence of families and schools”. *Education Next* 16(2), 22–30.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann (2022). “Patience, Risk-Taking, and Human Capital Investment across Countries”. *The Economic Journal* 132(646), 2290–2307.

Hanushek, Eric A and Margaret E Raymond (2005). “Does school accountability lead to improved student performance?” *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 24(2), 297–327.

Hanushek, Eric A, Jens Ruhose, and Ludger Woessmann (2017a). “Knowledge capital and aggregate income differences: Development accounting for US states”. *American Economic Journal: Macroeconomics* 9(4), 184–224.

Hanushek, Eric A, Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann (2015). “Returns to skills around the world: Evidence from PIAAC”. *European Economic Review* 73, 103–130.

– (2017b). “Coping with change: International differences in the returns to skills”. *Economics Letters* 153, 15–19.

Hanushek, Eric A and Ludger Woessmann (2008). “The role of cognitive skills in economic development”. *Journal of Economic Literature* 46(3), 607–668.

– (2011). “The economics of international differences in educational achievement”. *Handbook of the Economics of Education* 3, 89–200.

– (2012). “Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation”. *Journal of Economic Growth* 17, 267–321.

– (2015). *The knowledge capital of nations: Education and the Economics of growth*. MIT press.

Hartog, Joop and Luis Diaz-Serrano (2007). “Earnings risk and demand for higher education: A cross-section test for Spain”. *Journal of Applied Economics* 10(1), 1–28.

Hartog, Joop, Luis Diaz-Serrano, et al. (2014). “Schooling as a risky investment: A survey of theory and evidence”. *Foundations and Trends in Microeconomics* 9(3–4), 159–331.

Heckman, James J (1976). “A life-cycle model of earnings, learning, and consumption”. *Journal of Political Economy* 84(4, Part 2), S9–S44.

Henley, Amy J and Florence D DiGennaro Reed (2015). “Should you order the feedback sandwich? Efficacy of feedback sequence and timing”. *Journal of Organizational Behavior Management* 35(3-4), 321–335.

Hermes, Henning, Martin Huschens, Franz Rothlauf, and Daniel Schunk (2021). “Motivating low-achievers - Relative performance feedback in primary schools”. *Journal of Economic Behavior & Organization* 187, 45–59.

Hofstede, Geert (1991). *Cultures and organizations: Software of the mind*. Vol. 2. Mcgraw-hill London.

Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov (2010). *Cultures and organizations: Software of the mind*. Vol. 3. Mcgraw-hill New York.

Howitt, Peter and Philippe Aghion (1998). "Capital accumulation and innovation as complementary factors in long-run growth". *Journal of Economic Growth*, 111–130.

Hyde, Janet S., Sara M. Lindberg, Marcia C. Linn, Amy B. Ellis, and Caroline C. Williams (2008). "Gender Similarities Characterize Math Performance". *Science* 321(5888), 494–495.

Ichino, Andrea and Giovanni Maggi (2000). "Work environment and individual background: Explaining regional shirking differentials in a large Italian firm". *The Quarterly Journal of Economics* 115(3), 1057–1090.

Ilgen, Daniel R, Cynthia D Fisher, and M Susan Taylor (1979). "Consequences of individual feedback on behavior in organizations." *Journal of Applied Psychology* 64(4), 349.

ILO (2022). *Labour Force Statistics (LFS): Working-age population by sex and age (annual)*. Id: POP_XWAP_SEX_AGE_NB_A (accessed on 07 November 2022).

ISSP Research Group (2016). *International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012*. GESIS Datenarchiv, Köln. ZA5900 Datenfile Version 4.0.0, https://doi.org/10.4232/1.12661.

Jacobs, Marion, Alfred Jacobs, Margaret Gatz, and Todd Schaible (1973). "Credibility and desirability of positive and negative structured feedback in groups." *Journal of Consulting and Clinical Psychology* 40(2), 244.

Jung, Dawoon, Tushar Bharati, and Seungwoo Chin (2021). "Does education affect time preference? Evidence from Indonesia". *Economic Development and Cultural Change* 69(4), 1451–1499.

Kahn, Shulamit and Donna Ginther (2017). *Women and STEM*. NBER Working Papers 23525. National Bureau of Economic Research, Inc.

Kajitani, Shinya, Keiichi Morimoto, and Shiba Suzuki (2020). "Information feedback in relative grading: Evidence from a field experiment". *PloS one* 15(4), e0231548.

Kim, Hong-Kyun (2001). "Is there a crowding-out effect between school expenditure and mother's child care time?" *Economics of Education Review* 20(1), 71–80.

Kluger, Avraham N and Angelo DeNisi (1996). "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory." *Psychological Bulletin* 119(2), 254.

Koerselman, Kristian and Roope Uusitalo (2014). "The risk and return of human capital investments". *Labour Economics* 30, 154–163.

Kosse, Fabian and Friedhelm Pfeiffer (2012). "Impatience among preschool children and their mothers". *Economics Letters* 115(3), 493–495.

Lavecchia, Adam M, Heidi Liu, and Philip Oreopoulos (2016). "Behavioral economics of education: Progress and possibilities". *Handbook of the Economics of Education*. Vol. 5. Elsevier, 1–74.

Levhari, David and Yoram Weiss (1974). "The effect of risk on the investment in human capital". *The American Economic Review* 64(6), 950–963.

Lucas Jr, Robert E (1988). "On the mechanics of economic development". *Journal of Monetary Economics* 22(1), 3–42.

Mankiw, N Gregory, David Romer, and David N Weil (1992). "A contribution to the empirics of economic growth". *The Quarterly Journal of Economics* 107(2), 407–437.

Mayer, Thierry and Soledad Zignago (2011). "Notes on CEPII's distances measures: The GeoDist database".

Mendez, Ildefonso (2015). "The effect of the intergenerational transmission of noncognitive skills on student performance". *Economics of Education Review* 46, 78–97.

Miao, Bin and Songfa Zhong (2015). "Comment on "Risk preferences are not time preferences": Separating risk and time preference". *American Economic Review* 105(7), 2272–2286.

Mincer, Jacob (1958). "Investment in human capital and personal income distribution". *Journal of Political Economy* 66(4), 281–302.

– (1974). "Schooling, Experience, and Earnings." *NBER*.

Mischel, Walter, Yuichi Shoda, and Monica L Rodriguez (1989). "Delay of gratification in children". *Science* 244(4907), 933–938.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat (2022). "Managing self-confidence: Theory and experimental evidence". *Management Science* 68(11), 7793–7817.

Moffitt, Terrie E, Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent W Roberts, Stephen Ross, et al. (2011). "A gradient of childhood self-control predicts health, wealth, and public safety". *Proceedings of the National Academy of Sciences* 108(7), 2693–2698.

Moriconi, Simone and Núria Rodríguez-Planas (2021). *Gender Norms and the Motherhood Employment Gap*. IZA Discussion Papers 14898. Institute of Labor Economics (IZA).

Muis, Krista R, John Ranellucci, Gregory Trevors, and Melissa C Duffy (2015). "The effects of technology-mediated immediate feedback on kindergarten students' attitudes, emotions, engagement and learning outcomes during literacy skills development". *Learning and Instruction* 38, 1–13.

Murphy, Kevin M and Robert H Topel (1985). "Estimation and inference in two-step econometric models". *Journal of Business & Economic Statistics* 3(4), 370–379.

NEPS-Netzwerk (2021). *Nationales Bildungspanel, Scientific Use File der Startkohorte Studierende*. Tech. rep. Leibniz-Institut für Bildungsverläufe (LIfBi), Bamberg.

Obradovich, Nick, Ömer Özak, Ignacio Martıén, Ignacio Ortuño-Ortıén, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan, and Ángel Cuevas (2022). "Expanding the measurement of culture with a sample of two billion humans". *Journal of the Royal Society Interface* 19(190), 20220085.

OECD (2016a). *Technical report of the survey of adult skills (PIAAC), Second Edition*.

OECD (2016b). *The survey of adult skills: Reader's companion*. OECD Publishing.

– (2019). "PISA 2018 Results (Volume I)".

– (2020). *Harmonised unemployment rate (HUR) (indicator)*. doi: 10.1787/52570002-en (accessed on 07 November 2022).

Ortego-Marti, Victor (2017). "Differences in skill loss during unemployment across industries and occupations". *Economics Letters* 161(C), 31–33.

Oster, Emily (2019). "Unobservable selection and coefficient stability: Theory and evidence". *Journal of Business & Economic Statistics* 37(2), 187–204.

Pagan, Adrian (1984). "Econometric issues in the analysis of regressions with generated regressors". *International Economic Review*, 221–247.

Palacios-Huerta, Ignacio (2003). "An empirical analysis of the risk properties of human capital returns". *American Economic Review* 93(3), 948–964.

Petrongolo, Barbara and Maddalena Ronchi (2020). "Gender gaps and the structure of local labor markets". *Labour Economics* 64(C).

Potrafke, Niklas (2019). "Risk aversion, patience and intelligence: Evidence based on macro data". *Economics Letters* 178, 116–120.

Putnam, Robert D, Robert Leonardi, and Rafaella Y Nanetti (1992). *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.

Rebollo-Sanz, Yolanda F and Sara De la Rica (2020). "Gender gaps in skills and labor market outcomes: Evidence from the PIAAC". *Review of Economics of the Household*, 1–39.

Resnjanskij, Sven, Jens Ruhose, Simon Wiederhold, and Ludger Woessmann (2021). "Can Mentoring Alleviate Family Disadvantage in Adolscence? A Field Experiment to Improve Labor-Market Prospects". *CESifo Working Paper 8870*.

Robinson, Joseph Paul and Sarah Theule Lubienski (2011). "The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings". *American Educational Research Journal* 48(2), 268–302.

Romer, Paul M (1990). "Endogenous technological change". *Journal of Political Economy* 98(5, Part 2), S71–S102.

Rosen, Sherwin (1976). "A theory of life earnings". *Journal of Political Economy* 84(4, Part 2), S45–S67.

Rump, Markus, Wiebke Esdar, and Elke Wild (2017). "Individual differences in the effects of academic motivation on higher education students' intention to drop out". *European Journal of Higher Education* 7(4), 341–355.

SAO/NASA (2022). *Astrophysics Data System: Share of female authors in astrophysics*. Available at https://ui.adsabs.harvard.edu/help/api/ (accessed on 07 November 2022).

Schaible, Todd D and Alfred Jacobs (1975). "Feedback III: Sequence effects: Enhancement of feedback acceptance and group attractiveness by manipulation of the sequence and valence of feedback". *Small Group Behavior* 6(2), 151–173.

Schroeders, Ulrich, Oliver Wilhelm, and Gabriel Olaru (2016). “The influence of item sampling on sex differences in knowledge tests”. *Intelligence* 58(October 2019), 22–32.

Schultz, Theodore W (1961). “Investment in human capital”. *The American Economic Review* 51(1), 1–17.

Schumpeter, Joseph A (1912). *Theorie der wirtschaftlichen Entwicklung*. Duncker und Humblot.

Slowiak, Julie M and Areanna M Lakowske (2017). “The influence of feedback statement sequence and goals on task performance.” *Behavior Analysis: Research and Practice* 17(4), 357.

Solow, Robert M (1956). “A contribution to the theory of economic growth”. *The Quarterly Journal of Economics* 70(1), 65–94.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim (2022). “Patience and comparative development”. *The Review of Economic Studies* 89(5), 2806–2840.

Sutter, Matthias, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann (2013). “Impatience and uncertainty: Experimental decisions predict adolescents’ field behavior”. *American Economic Review* 103(1), 510–531.

Thorndike, Edward L (1913). “Educational psychology, Vol 1: The original nature of man.” *Teachers College*.

– (1927). “The law of effect”. *The American Journal of Psychology* 39(1/4), 212–222.

Thorson, Kjerstin, Kelley Cotter, Mel Medeiros, and Chankyung Pak (2021). “Algorithmic inference, political interest, and exposure to news and politics on Facebook”. *Information, Communication & Society* 24(2), 183–200.

Todd, Petra E and Kenneth I Wolpin (2003). “On the specification and estimation of the production function for cognitive achievement”. *The Economic Journal* 113(485), F3–F33.

Tyng, Chai M, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik (2017). “The influences of emotion on learning and memory”. *Frontiers in Psychology* 8, 1454.

Villeval, Marie Claire (2022). “The cognitive and motivational effects of performance feedback”. *Encyclopedia of Labor Studies*. Edward Elgar Publishers.

Weiss, Yoram (1972). “The risk element in occupational and educational choices”. *Journal of Political Economy* 80(6), 1203–1213.

Woessmann, Ludger (2010). “Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries”. *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 230(2), 234–270.

– (2016a). “The economic case for education”. *Education Economics* 24(1), 3–32.

– (2016b). “The importance of school systems: Evidence from international differences in student achievement”. *Journal of Economic Perspectives* 30(3), 3–32.

Zadra, Jonathan R and Gerald L Clore (2011). “Emotion and perception: The role of affective information”. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(6), 676–685.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez (2019). “When students don’t care: Reexamining international differences in achievement and student effort”. *Journal of Human Capital* 13(4), 519–552.

Zax, Jeffrey S and Daniel I Rees (2002). “IQ, academic performance, environment, and earnings”. *Review of Economics and Statistics* 84(4), 600–616.

Zimmermann, Florian (2020). “The dynamics of motivated beliefs”. *American Economic Review* 110(2), 337–61.