Charles University in Prague

Centre for Economic Research and Graduate Education -
Economic Institue

MASTER THESIS

Jiří Grosman, BSc

Role Models, High School Management and Female Engagement in STEM: Evidence from the Czech Republic

Supervisor of the Master Thesis: PhDr. Jan Zapal, Ph.D.

Study Programme: MA Economic Research

Prague 2023

Master's Thesis Proposal

CERGE
Charles University

Author's name and surname: Jiri Grosman

E-mail: jiri.grosman@cerge-ei.cz

Phone: +420 734 714 945

Supervisor's name: PhDr. Jan Zapal Ph.D.

Supervisor's email: jan.zapal@cerge-ei.cz

Notes: Please enter the information from the proposal to the Student Information System (SIS) and submit the proposal signed by yourself and by the supervisor to the Academic Director ("garant") of the undergraduate program.

Proposed Topic:

Role Models, High School Management and Female Engagement in STEM: Evidence from the Czech Republic

Preliminary scope of work:

Research question and motivation

The analysis of driving forces behind the gender attainment gap remains a subject undergoing intense study. Due to its impact on "labour-market outcomes (…) [and on] economic growth" (Smith & Sokolová, 2022), it is the educational attainment gap in particular that continues to draw the attention of scholars. Recent studies have shown the gap to be widening "in the direction of the girl" (Ibid.). Interestingly, however, the trend has not been accompanied by a decrease in the employment attainment gap, which traditionally favours the male. Women continue to be underpaid and underrepresented in top positions despite gains in education attainment. As has been shown, the wage gap "remains (…) time-invariant" (Zajíčková & Zajíček, 2021). The estimation of the effect structural labour market relations exert on women's gains to education is the subject of this study.

Contribution

While there is a plethora of evidence on either employment or educational attainment gaps (see Zajíčková & Zajíček, 2021), the literature on causal linkage between the two phenomena remains relatively modest in volume. This fact is unfortunate as the disentanglement of the processes behind the 'attainment shift' is an issue of policy importance. Crucially, the preferred solution of the decision-maker may be contingent on whether it is the case that boys' learning experience at school is differential relative to that of girls or whether it is the conditions in the labour market that perpetuate existing inequalities. Absent of uncovering such linkage, establishing whether to focus on differential learning outcomes or whether disemboweling structural sexism in the labour market ought to be a priority may be difficult. This paper strives to offer some clarity by evaluating the status quo and by making the relationship between the two phenomena explicit.

Methodology

To facilitate estimation, we need to construct a measure that allows us to separate the hypothesised causal effect of structural, gender-based disadvantage in the labour market from returns to education. While the conventional 'wage gap' to a certain extent captures such relationship, it represents a descriptive rather than a causal metric. That is what this study aims to remedy. The exact method used will ultimately be dependent on data availability. Merging the EU-SILC data (see Zajíčková & Zajíček, 2021) with CHPS data (see Smith & Sokolová, 2022) might present one viable option for constructing the counterfactual by employing the diff-in-diff method. Alternatively, the CLOSE dataset might also be used; these options, however, will have to be explored further.

Outline

I. Introduction
II. Theoretical Framework
III. Data
IV. Research Design
V. Results
VI. Discussion
VII. Conclusion

List of academic literature:

Bibliography

Ruhm, C. (1998). The Economic Consequences of Parental Leave Mandates: Lessons from Europe. The Quarterly Journal of Economics, Vol. 113(1): 285-317.
Sloane, C., Hurst, E. & Black, D. (2019) A Cross-Cohort Analysis of Human Capital Specialization and the College Gender Wage Gap. Becker Friedman Institute: Working Paper No. 121.
Smith, M. L. & Sokolová, V. (2022). Gender gaps in educational pathways in the Czech Republic. British Journal of Sociology of Education, Vol. 43(2): 296-313. DOI: 10.1080/01425692.2021.1971062.
Zajíčková, D. & Zajíček, M. (2021). Gender Pay Gap in the Czech Republic - Its Evolution and Main Drivers. Prague Economic Papers, Vol. 30(6): 675-723. DOI: 10.18267/j.pep.787
Zajíčková, D., Zajíček, M. & Rašticová, M. (2021). Does Anti-Discrimination Legislation Work? The Case of Motherhood Penalty in the Czech Republic. Employee Responsibilities

Date: 19.9.2022

_____
Guarantor

_____
Author

_____
Supervisor

Declaration of Authorship

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

On 31.7.2023 in Prague Jiří Grosman

Suggested Citation

GROSMAN, J. (2023) Role Models, High School Management and Female Engagement in STEM: Evidence from the Czech Republic. [Master's thesis, Centre for Economic Research and Graduate Education - Economics Institute, Charles University]. CU Digital Repository.

Length of the thesis: 121,705 characters (incl. spaces)

Disclaimer: Parts of this thesis were readapted from my personal works previously submitted to CERGE-EI over the course of my studies in the MA Economic Research study programme.

Title: Role Models, High School Management and Female Engagement in STEM: Evidence from the Czech Republic

Author: Jiří Grosman, BSc

Institute: CERGE-EI

Supervisor: PhDr. Jan Zapal, Ph.D., CERGE-EI

Abstract: The fact that the gender wage gap remains at least in part driven by women's self-selection out of profitable career choices has been an established fact in the labour economics literature. What has been given comparatively less attention by economists, however, are the motivating factors of such self-selection. In this thesis, I design an empirical strategy aimed to assess the strength of a particular driver of differential gendered choices: high school management, and the associated policy intervention. Specifically, I examine the STEM preferences of high school students using population data on Czech university applicants between the years of 2007-2021. By employing both matching and a difference-in-differences estimation framework to facilitate causal interpretation, I argue that school mentoring programmes, and the way the school is being run, can significantly affect the choices made by a non-negligible number of female students - an inference corroborated by the data and my subsequent analysis.

Keywords: *gender gap; STEM; high school; education*

Název práce: Význam lidských vzorů a řízení středních škol pro zapojení žen ve STEM oborech: analýza z prostředí České republiky

Autor: Jiří Grosman, BSc

Pracoviště: CERGE-EI

Vedoucí práce: PhDr. Jan Zápal, Ph.D., CERGE-EI

Abstrakt: Skutečnost, že *gender wage gap* je alespoň do určité míry motivován selekcí méně výdělečných profesí ze strany žen je v současném oboru ekonomie pracovního trhu v zásadě obecně přijímaným faktem. Čemu však byla doposud věnována srovnatelně menší pozornost odobrníků jsou faktory, jež tuto selekci způsobují. V rámci této diplomové práce si za pomoci vlastní výzkumné strategie kladu za cíl změřit důležitost jednoho specifického motivátoru rozdílných, gendrově podmíněných rozhodnutí: způsob vedení středních škol, a s ním spojených opatření. Konkrétně, skrze analýzu celorepublikových dat zachycujících údaje o přihláškách uchazečů o studium na vysokých školách z let 2007-2011 v této práci mapuji inklinaci studentů SŠ ucházet se o vědní (STEM) obory v dalším studiu. Za užití kauzálních *matching* a *difference-in-deifferences* modelů argumentuji, že školní mentoringové programy společně se způsobem, jakým je škola vedena, mohou statisticky významným způsobem ovlivnit studijně-profesní rozhodnutí nemalého množství studentek - tvrzení, v jehož prospěch hovoří jak mnou zmapovaná data, tak i má navazující analýza.

Klíčová slova: *gender gap; STEM; střední škola; vzdělávání;*

Acknowledgements

First and foremost, I wish to thank my mom, my grandmother, and all of my loving siblings for their enduring support throughout the process of completing this thesis. Second, I wish to express my gratitude to to the supervisor of this thesis, Jan Zápal, for being an open-minded and tolerant guide throughout the writing process. Third, a further thank you is in order to Martina Fucimanová of SÚ AV ČR for her kind provision of the mentoring treatment data, and to Tomáš Protivínský and Václav Korbel for their assistance in the cleaning and analysis of the *Uchazeč* datasets. Fourth, I also express my sincere gratitude to Daniel Münich, Štěpán Jurajda, and Sebastian Ottinger of CERGE-EI, who provided valuable insights in various stages of the writing process. Fifth, I am deeply indebted to Mr. Jiří Balhar of Balservis Entreprises for being a LaTeXmaster that he is, and to my dear friends and roommates of Sokolovská 50, whose endearing care and affection has helped me always to keep dispiriting thoughts at arms' length. Sixth and finally, I express my admiration and gratefulness to the lovely team of Kavárna Mlýnská for the countless espresso shots, lime sodas, *hermelíns* and smiles that ensured that my brain and stomach always weathered the storm of heavy intellectual strain. Thank you, all.

Contents

Introduction

In the field of labour economics, the conventional model of human capital accumulation predicts that educational attainment is positively associated with the earning capacity of the agent. However, recently uncovered evidence on female educational outcomes and the persistence of the gender pay gap seemingly contradicts the model's predictions. As empirical findings across the Western world document, female students continue to outperform their male counterparts, be it during high school years or in tertiary education attainment (see e.g., Schwab et al., 2017; Smith and Sokolova, 2022; c.f. Legewie and DiPrete, 2009). Despite these recent gains, in many countries - including the Czech Republic - the gender wage gap remains unchanged (Zajickova & Zajicek, 2021). Reconciling these two seemingly contradictory observations continue to present an intriguing and insightful area of research with wide societal, as well as policy implications.

An important strand of the gender gap literature has focused on the underrepresentation of women in positions of responsibility (Bell, 2005), wage discrimination at the workplace (Jurajda & Paligorova, 2009), or the effect of maternal leave on the wage differential (Anderson et al., 2002). These areas are, indeed, most relevant to the decomposition of gendered labour market inequalities, and they present us with a plethora of policy and research problems in their own right. Arguably, though, one additional area continues to be given comparatively less attention: schooling, and high school education in particular.

While the original aim of this thesis, as stated in the Proposal, was to analyse the causal determinants of the wage gap that operate within the labour market, the study of the associated literature directed my attention to factors that *precede* rather than follow women's entry into professional life. Having learned that the motherhood penalty and women's self-selection into low-pay jobs count among the most influential factors in the formation of contemporary gendered inequalities in remuneration (see e.g., Hedija, 2016; Petrongolo and Ronchi, 2020), I chose

ultimately to focus on the latter.

The educational setting, and especially that of a high school, shows considerable capacity to affect subsequent study and career outcomes. Notably, the incentive structures employed by educators and educational institutions represent an influential component in the choices made by individual students. For one thing, just how fair, informative, or competitive the system of rewarding students proves to be can significantly affect further study decisions (see e.g., Federičová, 2019; Lavy and Megalokonomou, 2019). Similarly, the gender or the ethnicity of teaching administrators, or whether institutional measures outside the classroom are enacted, are equally likely to bear influence on individual choices (Evans, 1992; Robb and Robb, 1999; Carrell et al., 2010).

With these observations in mind, the study of high-school determinants of individual study choices - and the subsequent self-selection of a notable proportion of women into low-pay or non-technical occupations - thus appears to be appealing both in terms of understanding the gender pay gap and finding an appropriate remedy. It is this consideration in particular, together with the fact that a previously understudied, novel dataset on Czech university applications became available, that motivated the focus of the present thesis and its research design.

Specifically, exploiting Czech Ministry of Education (MŠMT) *Uchazeč* population data on university applicants spanning years 2007, 2011, 2017 and 2021, I hypothesise that both school-level intervention promoting female STEM engagement and the socioeconomic characteristics of the region can causally influence individual decisions to pursue STEM study tracks. By combining the *Uchazeč* datasets with data from the Czech School Registry (MŠMT, 2023a), and municipality-level statistics and geospatial files on Czech administrative districts (ČSÚ, 2023a; ČÚZK, 2023), I obtain a high-school-level measure of the STEM attainment gap, as well as a number of municipality-level controlling variables. My treatment data, on the other hand, originates from the Institute of Sociology

at the Czech Academy of Sciences (SÚ AV ČR), which administered a STEM mentoring programme for female students at selected high-schools between the years of 2010 and 2017.

Conceptualising the mentoring programme as the intervention of interest, I capitalise on the panel character of the resulting population dataset by employing a matching, and a difference-in-differences estimation strategy, thus testing for the influence of the pro-STEM engagement treatment, and of changes in the socioeconomic characteristics of administrative districts, on the STEM attainment gap. As I argue, measuring the STEM attainment gap as the proportion of female applicants opting for a STEM study field in a given high-school, such strategy allows me to credibly estimate whether, and to what extent, intervention and socioeconomic conditions affect STEM engagement among women. Such inference, I maintain, allows me to assess whether these channels of influence represent a suitable area for policy work designed to eliminate the persistent inequalities in STEM representation, and whether school-level intervention presents a suitable avenue for promoting STEM engagement.

This thesis is structured as follows. Distinguishing between the Theoretical Part and the Analytical Part, Section I opens by reviewing the relevant scholarly literature on the gender gap and linking it to the issue of the STEM attainment gap in the context of the contemporary Czech Republic. Having established the framework for analysis, Section II then develops the research question aimed to address the lack of scholarly attention to school-level determinants of STEM attainment gap and, crucially, the associated policy treatment. This concludes the Theoretical Part.

The Analytical Part commences with a discussion of the data sources, describing the manipulation thereof and the data cleansing processes, in Section III. Section IV follows with an overview of the present empirical design, including regression specifications and the discussion of possible alternative designs. Results and their discussion are then presented in Sections V and VI respectively,

contextualising the main findings and examining their main limitations and contributions. I conclude with a brief summary of the thesis and its main findings, highlighting possible avenues for future research.

Theoretical Part

I Background Information

Few areas have recently been given as much scholarly attention in labour economics as the gender gap. As is apparent from the discussions among both the academia and the general public, the conceptualisation of this gap - commonly understood as an undesirable, socially relevant inequality between men and women - is multifaceted, and spans many areas of our social and economic lives. For instance, in their construction of *Global Gender Gap Index*, the 2017 Global Gender Gap Report differentiates between four "fundamental categories" of gender equality: *Economic Participation and Opportunity*, *Educational Attainment*, *Health and Survival*, and *Political Empowerment* (Schwab et al., 2017, pp. 4–5). As Schwab et al. (2017) note, however, not even this measure can fully capture the the phenomenon in its entirety.

Given the breadth of issues that the concept of *gender gap* encompasses, it may first be advisable to explain to the reader how the term is being used in this thesis, and to narrow down the conceptual scope of inquiry, thus linking the relevant literature to the work at hand and motivating the present research design. Before we proceed to the specification of the research question, the following few subsections are therefore devoted to the discussion of the *gender attainment* and the *gender pay gap*, specifying what I understand them to denote, how they relate to each other, and what scholarly knowledge regarding their relationship to education has been uncovered to this date.

I.A The Gender Pay Gap

Despite the fact that this thesis primarily focuses on STEM attainment gap, and gender pay gap should thus *prima facie* represent only a tangential interest, I consider it instructive to begin with a short discussion of the latter. Observable

economic and financial outcomes are what tends to be of interest to most economic studies in gender inequality, after all (see e.g., Bishu and Alkadry, 2017). Given the relative ease with which pay gap can be measured and interpreted, this focus is understandable; the two phenomena are nonetheless tightly intertwined, as the review that follows reveals. Discussing one without the other would therefore fail to adequately reflect the nuance in the mechanisms via which they materialise, as well as their mutual interaction.

The idea behind the notion of *gender gap*[1] broadly conceived is relatively simple. The Global Gender Gap Report defines it as "the average income earned by women, relative to income earned by men" (Schwab et al., 2017, p. 36), a definition that I choose also to follow for the sake of simplicity and conceptual clarity throughout this thesis.[2]

A useful overview of the most recent advancements in the gender pay gap literature has been given by Petrongolo and Ronchi (2020), whose decomposition of the pay gap into its motivating factors provides a useful indication as to the linkage between the gendered differences in *attainment*, and in *remuneration*. In their paper, Petrongolo and Ronchi (2020, p. 3) identify the most prominent determinants of the pay gap, comprising employment gap, 'vertical' underrepresentation of women in "high-income, high-status occupations," and differential preferences and attitudes to risk, competition, and negotiations of women relative to men. Arguably, each one of these factors corresponds to a distinct research area with its respective strand of literature; the following few paragraphs nonetheless offer at least a basic overview of the state of contemporary academic knowledge - in-

[1]Note that for the purposes of this paper, gender pay gap is meant to denote *raw gender pay gap* – i.e., the total difference in average per capita income between males and females. Other measures, e.g., ones controlling for parental leave, or comparing income of men and women at the same employment position, are also being used in the literature. Unless explicitly stated otherwise, these are not being considered here.

[2]Alternative definitions - e.g., ones based on *wage rates* rather than *income* - are also being used in the literature (see e.g., Hedija, 2016). Whether one is more preferable to the other arguably remains an open question; given that the primary focus of this thesis is on *attainment gap*, I evade this discussion by working with Schwab et al.'s 2017 conceptualisation absent of presenting as the only acceptable definition. Admittedly, working with an alternative, wage-based definition is indeed possible; with that being said, though, I deem any substantive effect of doing so on the discussion that follows to be unlikely.

formation which will prove especially useful in terms of understanding the role of education, and the associated STEM attainment gap, to the development of these processes.

First, turning to employment gap – i.e. the gendered difference in employment participation rates – as one of the main drivers of pay gap occurrence, a notable contribution in this area was made by Fitzenberger et al. (2004). Using West German 1976-1995 Mikrozensus data, the authors construct a longitudal model uncovering a convergent trend in gender-contingent participation rates. Despite this reassuring trend, however, inequalities still persist, "rang[ing] from 5 percentage point[] (…) to 30 percentage point[]" difference in participation rates across young and middle-aged age groups respectively Fitzenberger et al. (2004). These findings appear especially pertinent, as previous estimates from a study conducted by Anderson et al. (2002) show that frequent "[a]bsences from the labor market" are largely to blame for the so-called *motherhood penalty* - i.e., the fact that mothers tend to earn less on average relative to their childless counterparts (Anderson et al., 2002, p. 354). Moreover, Anderson et al.'s (2002) findings show that the penalty tends to be higher for high-skilled workers, a result of high significance to the educational attainment gap literature.

The vertical underrepresentation of women relative to men, on the other hand, has previously been studied by Bertrand and Hallock (2001, p. 17), who finds that across the US corporations 1990s, only a "small fraction" of 2.4 % executives were women. Moreover, Bertrand and Hallock (2001, p. 3) also estimate such women to earn "about 45 % less than men" in comparable positions - an effect whose size is roughly comparable to the estimates of a similar study conducted by Bell (2005). These findings were also confirmed in the context of the Czech Republic, which is of direct interest to this thesis, by Jurajda and Paligorova (2009). In this study, Jurajda and Paligorova (2009) use firm-level Information System of Average Earnings (ISAE) data spanning years 2000-2004 to employ the Oxaca-Blinder and matching-based decompositions to find "a gender wage gap of about

20 percent." Jurajda and Paligorova (2009, p. 26) Controlling for demographics, firm and hierarchical position, the authors maintain that this difference is mainly due to the underrepresentation of women in highest-paying companies - a further validation of Bertrand and Hallock's (2001) findings from the context of the US.

Finally, the 'differences in preferences and attitudes' strand comprises a rich field, with significant overlaps to sociology and other related disciplines.[3] Importantly for our purposes though, the 'differences in attitudes' causal chain is of direct relevance to the attainment gap, and thus to the present research design (see Section I.B below).

Specifically, one may wonder whether these 'differences in attitudes' - as discussed, for instance, by Marianne (2011), or, in the context of labour economics, Rousille (2023) - indeed represent a sort of 'innate' or 'natural' preference towards certain types of behaviour, or a natural reaction of rational agents to incentive structures and reward schemes that have previously been put in place. With many of our preferences and attitudes being developed during our formational years of elementary to higher education, it appears more than plausible that the type of school, the teachers that are employed by it, and the reward schemes they use will affect the subsequent study- and career-related preferences of the individual. This is indeed the underlying idea of a number of scholarly studies studying the relationship between educational experience, and the later individual outcomes. The most relevant ones of these papers that relate to the issue of attainment, and STEM attainment gap specifically are being discussed in the section that follows.

I.B The Gender Attainment Gap

As has been discussed in the previous subsection, a multitude of factors plays an empirically established role in the emergence of gender pay gap. One could

[3]For the discussion of relevant scholarly knowledge outside the field of economics, and the corresponding theories (notably the gendered socialisation theory, which is indeed relevant for our purposes), see Reinking and Martin (2018).

plausibly argue, however, that there is a common denominator to the aforementioned drivers of gender pay gap – education. As I argue above, it would not be unreasonable to maintain, for example, that formation could bear relevance to employment preferences, self-selection into non-technical or lower-paying jobs, group behaviours, etc. Indeed, as the research of other authors suggest, the period of childhood and early adulthood considerably influences labour market outcomes later in life (see e.g., Smith and Sokolova, 2022; Brenøea and Lundberg, 2018). Drawing on the predictions of the human capital accumulation model, we can expect education to be especially relevant.

One general trend that has been studied extensively in the recent years, and that has been reaffirmed by the recent Smith and Sokolova (2022) study in the Czech context, is that of a consistent attainment gap - i.e., the difference in the aggregate level and quality of attained education between the respective genders - favouring the female student. Measured by either high school and college completion rates, or the grades achieved over the course of their studies, this finding is especially relevant given the existing gender *pay* gaps which – despite this fact – remain roughly constant (Zajickova & Zajicek, 2021).

Moreover, although we observe a smaller, or, as in the case of the Czech Republic, the United States, and other the Western countries, a *reverse* attainment gap favouring the female (Schwab et al., 2017), this does not necessarily mean that the representation of women rises evenly across all disciplines. In fact, even among the most egalitarian countries such as Sweden, Slovenia or New Zealand, the STEM attainment gap in tertiary education favouring the *male* still persists (Schwab et al., 2017).

In this thesis, as in many other scholarly papers, the primary focus is that on *high-school* environment. Arguably, by virtue of substantively affecting tertiary education and career decisions absent of substantially restricting the individual's subsequent choices, high-school education specifically presents a particularly suitable environment to study the effects of changes in educational setting on labour

market outcomes. It is presumably also this thought that led to the emergence of a particular strand of the education attainment gap literature that studies the relationship between the environment of a high school, and various other educational as well as economic metrics.[4] A particular sub-species of this literature strand is the STEM attainment gap, the review of which is given in the following subsection.

Before we move on to the detailed discussion of this phenomenon, though, a final note on attainment gap more generally is in order. The presumption that high-school education can be influential to subsequent study or career decisions is, arguably, far from being the only reason as to why both gender gap researchers and policymakers should be interested in studying educational settings. If, as one could maintain, high school education disproportionately motivated one gender group to take up certain types of study tracks and careers (which a number of scholarly findings in Section I.C indeed suggest), this could result in an over-representation of such group in certain occupational segments - a pattern that we observe in virtually all countries to this day, as the discussion above attempts show.[5] Beyond a certain threshold, this can present a concern, potentially leading to the emergence of mostly 'female' and mostly 'male' occupations.

Apart from being problematic from the point of stereotypisation of men's and women's career tracks, such occupational segregation is, as Hegewisch et al. (2010, p. 1) points out, both "inefficient economically", preventing people from working in areas that they could excel in, and a "major cause" of the persistent pay gap. Studying educational settings, and their relationship to subsequent, gender-contingent study and career decisions, is therefore relevant not only from the perspective of the appraisal of educational systems and their objectives, but also in terms of economic efficacy and gender gap elimination.

[4]For a systematic review of the extant literature on this (general) topic, see Verdugo-Castro et al. (2022).

[5]For additional details regarding this observation, I invite the reader to consult Part 2 of the Global Gender Gap Report (Schwab et al., 2017), which provides a country-by-country breakdown of the most relevant gender gap statistics, together with the associated contextual data.

I.C STEM and the Attainment Gap

This subsection discusses a particular strand of the attainment gap research which is of direct relevance to the present research design - i.e., the *STEM attainment gap* literature. As the term suggests, the papers in this strand focus primarily on the stark imbalances between the representation of men and women in Science, Technology, Engineering, and Mathematics (STEM) occupations and study. Studying the STEM field is appealing on multiple grounds, notably: the presence of stark imbalances in representation, comparatively high degree of in-field wage equality, and the role of education in pre-determining subsequent study tracks. Let me address each one of these areas in turn.

Imbalances in Representation

For one thing, the vast differences in STEM employment rates for males and females – as documented by Schwab et al. (2017), Moss-Racusin et al. (2018), Mouganie and Wang (2020), and other authors – are indeed intriguing. This is especially so considering the fact that despite the recent advances in female tertiary education attainment (see e.g., Smith and Sokolova, 2022), these gains have failed to fully materialise in the field of STEM. In fact, for some STEM fields and occupational areas, the progress has "completely stalled", or has even reversed, as Hegewisch et al. (2010, p. 1) remarks. Building on the discussion developed in Section I.B, this fact arguably begs the following question: is it really that women could have such an innate dislike of STEM-related disciplines and occupations, or is the presence of gendered discrepancies attributable to the way our educational system and/or the institutions are being run? It is the hope that the results uncovered in this thesis will also serve as supplementary evidence on this matter.

STEM Field and Gender Equality

Second, as some scholarly findings suggests, it could be maintained that within the field of STEM, the wage differential tends to be lower in STEM-related professions relative to other occupational areas. This is indeed what Kuhn and Shen (2013) observe in their analysis of China's *Zhaopin.com* job advertisement data, finding a negative relationship between wage discrimination and the skill requirement. This result is encouraging, as if this finding were to hold in other settings as well, the elimination of the STEM attainment gap could prove helpful in the objective of ameliorating the gender pay gap, too. More importantly (and somewhat counterintuitively) still, Kuhn and Shen (2013) arrives at this result despite the previously noted fact that in STEM study and occupational fields, women tend to be severely underrepresented (Mouganie & Wang, 2020).

Naturally, it needs to be stressed, however, that Kuhn and Shen's (2013) inference should be interpreted in the wider context of other scholarly findings. For one thing, it is questionable whether the estimates of the reduced wage differential are transferable to contexts outside China, and to public sector or academia. As Vohlídalová (2021) points out, for instance, in the context of Czech research institutions, this certainly does not appear to be the case. On the contrary, in Czech academic environment, it is STEM disciplines that exhibit the highest wage differential - despite the fact that "disproportionately more money goes to these fields." (Vohlídalová, 2021, p. 171)

One should similarly note that Kuhn and Shen's (2013) findings do not remotely imply that the presence of discriminatory practices does not constitute a deterrent to the entry into the field. Indeed, as a recent experimental study conducted by Moss-Racusin et al. (2018, p. 651) documents, perceived gender bias in research departments is likely to causally affect women's trust, "aspirations to participate in STEM", and the "sense of belonging" vis-á-vis STEM-oriented institutions.[6] Combined with the fact that there are very few female STEM grad-

[6]In a similar vein, the importance of labour market characteristics, and the degree to which it is marred by the lack of "gender equity", is found by Gevrek et al. (2020, p. 19) to be a

uates to apply for STEM research positions to begin with, the resultant STEM attainment gap is, after all, not that surprising a result.

STEM Attainment and Education

This final observation should rightly refocus our attention to the third and final area, as originally identified at the onset of this section: the relationship between the gendered differences in STEM, and education. Admittedly, the set of factors capable to manipulate STEM attainment gap that were previously analysed by the researchers is relatively rich. I therefore list only a limited selection of the ones most pertinent to high school education, and to the research question developed below.

Considering the frequency and intensity of in-class interaction between fellow students, one factor worthy of consideration here is that of *peer effects*. Admittedly, this specific channel is also particularly relevant to this thesis, as the main hypothesis of my research design in part relies on the presence of in-class spillover effects between students (see Section II.A for details). The idea behind the 'peer effects channel' is fairly straightforward - if students from an under-represented group see their fellow in-group classmate excel in a given field, they will be more likely to pursue the given study track themselves. The explanation of this linkage relies on the "sociopsychological" interpretation of in-class interaction: as Mouganie and Wang (2020, pp. 835–836) explain, the presence of "high-performing peers of the same gender may update girls' beliefs about their own mathematical ability, [thus] mitigating the effects of negative gender stereotypes."

This is indeed a result obtained by the two authors through their analysis of student-level administrative data for cohorts graduating between years 2007-2010

statistically significant predictor of female engagement in maths. While Gevrek et al. (2020) admittedly employs a cross-sectional and panel data techniques, rather than conducting an experiment akin to that of Moss-Racusin et al.'s (2018), Gevrek et al.'s (2020, p. 1) findings appear to lend further credence to the role of "societal gender equity" in the reduction of STEM attainment gap.

in China. Exploiting within-school variation in class composition, Mouganie and Wang (2020) identify the relationship between high-performing female peers and other students' STEM take-up. In doing so, the authors demonstrate the presence of an "affirmation effect", a fact of direct relevance to school policymakers Mouganie and Wang (2020, p. 807). Whether the authors' results are externally valid and can be extended to contemporary Western contexts as well, however, remains an open question, especially given the fact that a previous US study conducted by Evans (1992, p. 209) found "no evidence of a gender role-model effect."[7]

Applying a similar logic, the importance of the teaching staff and its characteristics on (female) student behaviour has also been documented in a number of scholarly studies. Examining the choice of major for junior college students in Texas, US, Porter and Serra (2020, p. 226) experimentally manipulates exposure of female students to successful female economists, finding a "large", role-model effect on "enrollment in further economics classes." Carrell et al. (2010), on the other hand, studied role-model effects in the context of STEM specifically, documenting the importance of the instructor's gender on female performance in STEM subjects. Working with individualised student and professor data at the the U.S. Air Force Academy (USAFA), the authors exploit a system of random course allocation of students to causally identify a positive relationship between STEM performance and being taught by a woman for female students.

This inference, again, is of direct relevance to the prevalence of STEM attainment gap, and demonstrates that the decisions made at the institutional level - be it the composition of staff, policies of encouragement, or reward schemes (see below) - can indeed bear substantial influence on the persistence of gender attainment gap. Here again, however, concerns about external validity are very much in order given the specific setting of the USAFA, and American military

[7]Although Evans (1992) did find a statistically significant role-model effect for African-American student subgroup, thus falling short of invalidating the translational channel of peer-effects *in general*.

education more generally. Admittedly, some studies - such as the one analysing microeconomics students at Brock University in Ontario, Canada by Robb and Robb, 1999 - fail to uncover a role-model effect based on gender of the instructor. Such studies, however, do not generally deny the existence of the mechanism - as Robb and Robb (1999, p. 17) note, "there may be a role-model effect that we could not find because of the particular set of female instructors looking," encouraging other researchers to study the phenomenon of teacher role-models further. It is the hope of this thesis to make a contribution to this strand of literature also.

Finally, direct encouragement, rewards and incentive structures can also be of direct relevance to subsequent study choices. Specifically, as Mechtenberg (2009) theorises by constructing a simple economic model of student-professor in-class interaction, imbalanced teaching practices, such as gendered grading bias, are likely to significantly affect educational outcomes - an effect that tends to be largest for the highly educated in terms of the resultant wage differential (Mechtenberg, 2009).[8]

Such findings have recently been validated empirically by a study conducted by Lavy and Megalokonomou (2019) in the context of Greek high schools. In a panel setting spanning eight years and more than 400 teachers in 21 schools, Lavy and Megalokonomou (2019) finds a statistically significant, negative relationship between gender bias favouring the male, and school attendance, university admission exam scores, and - crucially - the the field of subsequent study for the female student. An intelligible and easy-to-follow explanation of this relationship, as well as supplementary empirical evidence on this matter, have, then, been presented in a recent Czech study conducted by Federičová (2019).

Conceptualising grading bias as a "biased signal", Federičová (2019, p. 23) maintains that grading outcomes that are marred by teachers' perceptions of stu-

[8]Note that this result is also more or less in line with the empirical findings of Anderson et al. (2002), as discussed in Section I.A, in the sense that it is the high-skilled individuals that end up being proportionally more severely penalised, as measured by the gender pay gap following their entry into the labour market.

dents' non-cognitive skills rather than actual performance can lead to inaccurate updating of beliefs regarding one's own abilities on the part of the student. If this is the case, Federičová (2019, p. 2) claims, then the students' "further educational ambitions" can be "distorted", and their subsequent educational and professional tracks "substantially influence[d]" - an argument well in line with the previous empirical findings of Lavy and Megalokonomou (2019), and other authors.

This set of findings, again, highlights the importance of educational institutions to the attainment gap, albeit from a slightly different, more institutionally-relevant angle than the studies previewed at the onset of this section. Arguably, though, in-class incentive structures are likely comprise only a part of the educational mechanisms by which STEM attainment gap emerges. For instance, as Protivínskỳ and Münich (2018) document in the context of Czech high schools, the grading bias in mathematics and native language favours the *girl*, rather than the boy. Given Lavy and Megalokonomou's (2019) findings, we would presumably then expect to observe a reduced attainment gap in these disciplines, which - certainly for the case of mathematics - is not the case in the Czech republic (see e.g., Smith and Sokolova, 2022; Schwab et al., 2017).

As is apparent from the review presented above, the issue of STEM attainment gap occurrence is too complex an issue, and cannot thus be reduced to a single explanatory factor. That is not to say, however, that in a given educational setting, one channel cannot be more relevant than the other. Building on this observation, in this thesis I propose an estimation strategy that recognises this complexity but, at the same time, attempts to synthesise these competing mechanisms by looking at school-level, rather than individual outcomes. It is the hope of the author that these results will provide a more policy-oriented insight into the issue of STEM attainment gap, and the potential measures aimed to ameliorate it. The details of the research question that motivated this empirical strategy is presented in the following section.

II Research Question

In the previous section, I identify a number of channels which have been found to be most relevant vis-á-vis STEM attainment gap by previous researchers. Majority of the studies discussed above, however, analyse each of these channels in isolation, or in settings which exploit only conventional data sources (e.g., PISA scores, national test scores data, etc.). Moreover, as the discussion on the role-model and the peer-effect channel reveals, some of the empirical evidence is difficult to reconcile.

Focusing on the Czech context in which this thesis is set, Smith and Sokolova (2022) and others have found the girl to outperform the boy in high school attainment, but when it comes to STEM application rates, though, the male students continue to dominate.[9] As the review above shows, presence of other, previously identified determinants of STEM attainment is sure to play the role. According to Robb and Robb (1999, p. 17), "women [can be] discouraged (...) by the style of their male teachers." Drawing on Moss-Racusin et al. (2018), the way the *institution* itself is being perceived can affect the willingness of underrepresented groups to participate. The grading biases, and the gender composition of the teaching staff, are also likely to enter the equation (Lavy and Megalokonomou, 2019; Carrell et al., 2010; Porter and Serra, 2020). Arguably, any single of these factors constitute a potentially influential channel of their own. It would nonetheless be a mistake to consider each one of them individually.

In an attempt to address these shortcomings, in this thesis I intend to take a different approach. Admittedly, the role of education has been central to a notable portion of the articles reviewed above; it is my belief, however, that the role of educational institutions *per se* has to an extent been neglected. In the empirical design employed in this thesis, it is the hope of the authors that at least some of these channels can be jointly estimated, and therefore synthesised

[9]Moreover, this is despite the findings of Protivínskỳ and Münich (2018), who document that in the field of mathematics, Czech republic high school environment shows a grading bias favouring the *girl*. For details, see Section I.C.

for policy purposes.

The motivation behind the attempt to study the joint effect of these channels by looking at school-level STEM attainment outcomes is driven by the emphasis on policy action, as advanced in this thesis. Arguably, be it curriculum, grading practices, balanced gender composition of the staff, or other (e.g., extra-curricular) school policies, virtually all of these areas fall within the remit of the management of an individual high school.[10][11] Moreover, given the granularity of Czech high school management, which tend to be established mostly by the regional, rather than the central government, and which gives considerable autonomy to individual school directors, treating each individual high school as a distinct, self-governing unit appears especially pertinent in Czech context.

Focusing on school-level intervention rather than, say, in-class measures is advantageous from the perspective of perceived institutional bias. Building once more on Moss-Racusin et al.'s (2018) contributions, it would be sensible to hypothesise that if the school is perceived to respect and encourage women's effort to break through into the fields which are being viewed as predominantly male (which STEM certainly appears to be), this is likely to bear positive influence. In this sense, the research focus of this thesis can arguably also appear appealing.

With these observations in mind, the following research questions appear to present a natural corollary to the pool of the topical scholarly knowledge discussed in the previous sections. First, to what extent can the manner in which a given high-school is governed can affect individual preference regarding the subsequent study? Second, can it effectively influence *gendered* preferences vis-á-vis tertiary educational and decisions via its internal measures and policies? Third, and finally, what sort of measures can be enacted to curtail the persistent STEM attainment gap in tertiary education?

[10]These nonetheless generally have to be governed in accordance with the guidance of the Ministry of Education, Youth and Sports (MŠMT), and the Law for Schooling, or in Czech, *Zákon č. 561/2004 Sb., o předškolním, základním, středním, vyšším odborném a jiném vzdělávání (školský zákon)*.

[11]In most instances, these tend to be the responsibilities of the School Director, whose execution of these duties is being overseen by the relevant School Board.

While it is well beyond the scope of this thesis to present a comprehensive answer to all of these questions, it is my belief that the research design employed in this work may in the very least allow us to shed some additional light onto the latter two issues. Specifically, using a natural experiment in the form of a mentoring programme administered by the Institute of Sociology of Czech Academy of Sciences (SÚ AV ČR) in a number of Czech high schools, this thesis hopes to provide an additional empirical insight onto the role-model and peer-effects channels, which arguably present a crucial component to any successful institutional intervention.

Having presented the core of the research agenda of this work, and positioned it within the framework of contemporary scholarly knowledge, I am now in a position to formulate the main testable hypotheses, thus laying the ground for the introduction of my empirical design in the Analytical Part that follows. The introduction and the motivation of these hypotheses is the subject of the following two subsections, which also conclude the Theoretical Part of this thesis.

II.A The Intervention Hypothesis

As has been suggested, in their combined force, the joint influence of stereotypisation, sorting into 'male' and 'female' occupations, educators' behaviour, and peer effects can substantially affect the choices made by women (*and* men) regarding their preferred academic and professional tracks. In line with Federičová's (2019, p. 23) contribution, these influences, too, can be conceptualised as "biased signals" in the sense that they convey an inaccurate message to the individual regarding their cognitive skills, the potential to develop them, or to become successful applying them in the field. Even if we evaluate such 'biasedness' simply through the lens of economic utility, one would have to infer that this is likely to lead, to put it in Hegewisch et al.'s (2010 terms, 'inefficiencies'.

Measures and policy tools that can correct these biased perceptions of oneself or of one's abilities are therefore worth exploring further. As the results from the

literature suggests, even a small and inexpensive measure can sometimes suffice for an individual to update their beliefs, and to adjust their behaviour accordingly (see e.g., Bedard et al., 2021).

Turning to STEM engagement at the level of a high-school specifically, if, for instance, a school engages in an encouragement policy (e.g., through a STEM mentoring programme such as the one considered in this thesis), this could potentially go a long way in eliminating the persistent stereotypical perceptions. Consequently, by virtue of updating the commonly held beliefs regarding, say, the female students' purported 'natural' ineptitude for STEM, or by learning that female peers or women professionals can excel in a male-dominated field, the propensity of female students to explore a STEM study or professional track is likely to increase, *ceteris paribus*. Applying this logic, I proceed to the formulation of the main hypothesis of this thesis:

Hypothesis 1. High schools that engaged in a female STEM encouragement intervention will exhibit higher female STEM application rates for tertiary education.

Naturally, although this hypothesis is of direct relevance to the research question of this Theses, STEM encouragement interventions are nonetheless unlikely to fully explain the presence of differential STEM attainment patterns across schools or time. Instead, a substantial part of the heterogeneity the school-by-school STEM application rates is likely to come from medium- to long-term determinants of the socioeconomic environment in which the schools, and consequently the students, are set. This idea is being developed, and tested for, by the supplementary, Socioeconomic Hypotheses, the formulation of which can be found below.

II.B The Socioeconomic Hypotheses

Besides the main hypothesised relationship between pro-engagement STEM measures and women STEM attainment, a range of scholarly findings also hint at sub-

stantial, quantifiable differences in educational outcomes between the socioeconomically advantaged, and socioeconomically disadvantaged regions. As Prokop et al. (2021) documents, educational inequality appears to be an extremely relevant issue to the Czech Republic.

According to Prokop et al.'s (2021, pp. 84-85) findings, the region in which the student and their family resides, together with the opportunities for development it provides them, can "substantially ameliorate the future life path of the child." Moreover, as additional empirical findings suggest, the differences between the peripheral and the metropolitan municipalities appear to be especially pronounced, be it from the perspective of educational opportunities, in terms of the quality of life, or sociopolitical attitudes (see e.g., Murgaš and Klobučník, 2016; Dvořák et al., 2022). It is this regional dynamic that I hope to capture in my formulation of the following Socioeconomic Hypothesis:

Hypothesis 2. High schools that are located in metropolitan municipalities will exhibit higher female STEM application rates.

The motivation for the formulation of this hypothesis is as follows. Even though Prokop et al.'s (2021) analysis pays only limited attention to the issue of gendered differences in attainment, it appears sensible that the socioeconomic and cultural differences between the rural and the metropolitan areas should play a role. Affecting individual preference through expanded access to social and cultural resources, resources for personal development, and a richer set of economic opportunities, the decisions made by individuals living in metropolitan or economically affluent areas are likely to substantively diverge from the rest of the country.

As I maintain, there is no reason to believe why these should not extend to gendered study decisions as well. Moreover, one could also plausibly argue that peripheral and/or socioeconomically disadvantaged municipalities will also have a more restricted access to resources which question existing stereotypes, and which present both men and women as capable to excel in occupations that

are not 'traditionally' ascribed to them. With these observations in mind, my formulation of the hypothesis above can, then, be viewed as advantageous in the sense that it allows us to assess the presence of these mechanisms quantitatively.

Having introduced both my main and my supplementary hypothesis, I am now in a position to conclude the Theoretical Part of this thesis, moving on to the specifics of my research design aimed to test for these relationships empirically. That is the subject of the Analytical Part that follows.

Analytical Part

Over the course of the previous part of this thesis, I have provided the reader with an overview of the existing scholarly findings, establishing my research within the broader framework of the gender attainment gap research. Having identified the gaps in the contemporary knowledge that this research effort hopes to seal, in this part I move away from the domain of theory toward application. By employing the empirical design presented in the subsequent two sections, which to a large extent relies on difference-in-differences and matching identification strategies, I aim to provide a quantitative assessment of the hypotheses as outlined in Section II.

In my analysis, I proceed as follows. I begin by introducing the relevant data sources in Section III, providing an overview of the data cleansing and transformation processes fundamental to the creation of a reliable, school-level measure of the female STEM attainment gap. I follow with a brief overview of the empirical framework for both of my main, difference-in-differences, and matching specifications, as well as a supplementary pooling exercise. I, then, present an overview of the main results, comprising general time trends and spatial patterns in Czech STEM attainment and the main findings of the regression models presented in Section IV. Section VI concludes by a discussion of the results, outlining the chief contributions and limitations of the presented empirical framework.

III Data

As suggested in the Introduction, the exploratory, data-oriented focus of my research design arguably presents one of the main advantages of this thesis. In this section, I attempt to motivate such inference by presenting the requisite data sources, and the notable steps in my manipulation thereof.

Specifically, the sources exploited in this thesis comprise the following: the Czech Ministry of Education, Youth and Sports' (MŠMT) *Uchazeč* population

data on Czech university applications; the mentoring programme treatment data from the Institute of Sociology at Czech Academy of Sciences' (SÚ AV ČR); STEM classification data computed on the basis of MŠMT *Performance statistics of public and private colleges and universities* (MŠMT, 2023b); high school records from the MŠMT *School and school facilities registry* (MŠMT, 2023a); and the municipality-level controls of the Czech Statistical Office's (ČSÚ) *Municipality and district statistics* (MOS) database (ČSÚ, 2023a). The complementary geospatial files were, then, obtained from the Czech Office for Surveying, Mapping and Cadastre's (ČÚZK) RÚIAN database (ČÚZK, 2023).

Before we proceed to the discussion of these data sources, however, a quick sidenote is in order. As is the norm in comparable empirical work, the reader should be aware that only the data manipulation steps that are conventionally reported or that could potentially bear an influence on the outcomes of the analysis have been detailed in this section. Ideally, these should fully capture the core of the data preparation processes, the remainder of which can plausibly be viewed as perfectly non-controversial. For the sake of transparency, and to dispel any potential doubts regarding my approach, I nonetheless invite any interested reader to consult the associated scraping, cleansing and formatting *R* code, as presented in Appendix II. The matter of data manipulation transparency being off the table, let us move to the discussion of the main data sources of this work.

III.A *Uchazeč* Application Dataset

The main data source exploited in my analysis is the *Uchazeč* application dataset, comprising anonymised, individual-level data on all university or college[12] applications in the given year. The periods that were made available to the author were the years of 2007, 2011, 2017 and 2021. Of these years, the former two periods (i.e., 2007 and 2011) were previously uncleaned and unstudied; the latter two (i.e., 2017 and 2021), on the other hand, were already pre-processed for the

[12]Or, *vysoké školy* (VŠ) in Czech.

purposes of a forthcoming Protivínský and Korbel (2023) study.

By virtue of the necessity to align my data cleansing processes with those of Protivínský and Korbel (2023), which arose from the fact that the 2017 and 2021 data were available only in the pre-processed form, this setting proved unexpectedly advantageous. Although the stringency of filtering practices employed by Protivínský and Korbel (2023) arguably reduces the flexibility in the scope for my own analysis (see Figure 1 for details), my replication of Protivínský and Korbel's (2023) manipulation processes allows for a cross-validation of both my, and the other authors' resultant datasets. Moreover, as the discussion below shows, virtually all of the steps taken by Protivínský and Korbel (2023) are well in line with the objectives of the empirical design of my own. Consequently, the reduction in flexibility for the present thesis does not represent a significant obstacle to my research objectives.

The Original Dataset

The original, uncleansed 2007 and 2011 *Uchazeč* dataset of total 347,370 and 366,476 applications respectively covered a full breadth of information on the persona of the applicant and the admission process, encompassing the admission result, the form of the sought study, the year of the applicant's high school graduation, the area of their residence, etc. Crucially for my purposes, the listed variables also included the school registry number for the high school of the applicant's origin (i.e., the *IZO* code), the classification of the high school study programme (*obss*), the ID of the programme and the faculty that the individual applied to, and - importantly - a coded applicant ID number conveying the information about the gender of the applicant.

Already in this exploratory stage of the analysis, a number of difficulties came about. With the ultimate aim of constructing a school-by-school STEM application rate measure for both male and female students, and linking it with the treatment and the district- and municipality-level socioeconomic controls, addi-

tional information on the location and the type of the high school needed to be obtained. Similarly, considering the need to classify each individual application as STEM or non-STEM - a necessary step for the construction of the STEM attainment rate - an additional set of identifiers of the applicants' chosen programme and faculty had to be gathered.

School Identification Data

The need to link the *Uchazeč* dataset with additional high school and university information necessitated the acquisition of the following two resources: the MŠMT school registry data (MŠMT, 2023a), spanning the high school name, type and location, and the MŠMT university and college performance statistics MŠMT (2023b). Unfortunately, both of these resources were only available through an online portal with the possibility of displaying only a limited number of data entries.

Given the necessity to obtain the full dataset for either resource, I therefore decided to make use of a set of data scraping methods, employing the *rvest* and *RSelenium* packages in particular so as to obtain the requisite files. The specifics of the data gathering process are arguably too technical so as to be detailed here, the associated *R* code for the school registry data scrape can nonetheless be found in Appendix II. As per the university and college (VŠ) codes, these were originally obtained as a part of a research assistantship project under the guidance of prof. Ing. Štěpán Jurajda, Ph.D. at CERGE-EI. The associated *R* code for this second scraping exercise does not consist a part of this thesis, and will be provided by the author upon request.

Data Manipulation

Having obtained the identifying variables for both high schools and universities, I then follow Protivínský and Korbel's (2023) approach with the aim to ensure that the resultant *Uchazeč* dataset is free of any input errors, that unreliable

observations are removed, and that the final data comprise only of entries relevant to the research question at hand. This process comprises five major steps which are detailed in the paragraphs below, and they consist of filtering the data by the correct format of the ID code, the Czech citizenship, the graduation history, and the application result and the high school of origin being shown. A visual aid illustrating this process is presented in Figure 1.

First, I check for the format of the applicant ID code. Working as an indication of the quality of the given observation, the entries that consist of less than nine digits or that include special characters are likely to be invalid. Additionally, these observations also frequently suffer from significant missingness in other variables. Moreover, as the applicant ID includes the coded information on the applicant's gender, the fact that the format of the entry is incorrect prevents us from reliably extrapolating this information from the. This fact in particular motivated their removal, an approach in line with that of Protivínský and Korbel's (2023).

Reassuringly, for the year 2007 only 0.02% of observations (i.e., individual applications) had ID numbers of below 9 digits. Concurrently, for 2011 this proportion was considerably larger: as much as 1.29% of observations had an ID of length 8 or less. These proportions can nonetheless be viewed as negligible, and given that further checks showed no systemic relationship to other observables[13], their removal is unlikely to affect the subsequent analysis.

Turning to filtering by citizenship, in accordance with Protivínský and Korbel (2023) I further filter by the Czech nationality of applicant. While, admittedly, this step is not of direct relevance to my empirical setting, focusing on Czech applicants allows us to evade the potential issue of the effect being driven by non-domestic applicants, which could reflect dynamics that are different to the ones I formulate in my hypothesis. Nonetheless, given the fact that non-native entries also exhibit significant missingness in other variables and are, too, much more likely to show input errors, Protivínský and Korbel's (2023) decision to

[13]The exception to this fact is the dependence of shorter IDs on applicant nationality; since we are also filtering by Czech citizenship in the step that follows, this does not present a concern.

exclude them appears advantageous in this particular setting.

In any case, the proportions of non-Czech applications is relatively modest in the uncleaned data, spanning only 6.81% of observations for the 2007 dataset, climbing up to 9.93% for 2011. For Protivínský and Korbel's (2023) 2017 and 2021, these proportions are slightly higher, reaching 15.9% and 18.7% respectively (Protivínský & Korbel, 2023).

The applications are, then, further filtered by the criterion of recent graduation. Unless the applicant passed their high school completion exams [14] during the year leading up to the time of observation, the observation is excluded. In my setting, as in the setting of Protivínský and Korbel (2023), this appears sensible, as by virtue of excluding older graduates we are eliminating the possibility that any potential results will be driven by historical graduates' application decisions. More importantly still, by filtering by *Maturita* completion year, we are effectively eliminating continuing master's and doctoral applications - an important step in the identification of *high-school*, rather than *tertiary education* effects on subsequent study and professional decisions.

Quantifying the number of kept observations, the proportion of 2007 applications originating from applicants that passed their *Maturita* exam in 2007 was 58.67%, whereas for the year 2011, the respective proportion reached a similar level of 58.12% - a level largely driven by the fact that in this filtering exercise, we are effectively excluding continuing master's and doctoral applications.

Finally, Protivínský and Korbel (2023) filter the dataset by the result of admission being reported for the given application, excluding the observations for which the application outcome is not being shown. While, again, this is not of direct relevance to my empirical design, conducting this exercise serves as an additional check on the quality of the *Uchazeč* dataset. Considering the fact that for the 2007 dataset a mere 0.04% fail to show the admission result, and a total 0 of observations does so in the 2011 file, this exclusion should bear very lit-

[14] Or, *Maturitní zkouška/Maturita* in Czech.

tle influence. As the objective of my analysis is the construction of a school-level measure, one further condition needs to be met is that of the identifier of the high school of origin being displayed. While this is not an issue for the years covered by Protivínský and Korbel (2023)[15], in my 2007 and 2011 datasets, 1.67% and 1.06% of all observations respectively are characterised by a missing IZO variable. Unfortunately, since these observations, too, exhibit considerable missingness in other entries, it is not possible to conduct a balance check against those observations that are kept in this step; given the relatively low proportion of observations thus excluded, I consider this unlikely to have a detectable influence.

Effective Sample

Admittedly, the filtering process as outlined above could give the impression of being somewhat restrictive. As I have attempted to argue, however, the possible influences of the cleaning appear to be for the most part orthogonal to the effects that I am most interested in throughout my analysis. More compellingly still, the missingness appears to exhibit considerable overlaps across the variables which have been used for filtering, resulting in comparatively less observations being filtered out overall.

The resulting effective sample for years 2007 and 2011, from which the school-by-school STEM application ratios can be computed, therefore totals 191,174 and 195,518 application respectively, comprising 55.03% and 53.35% of the overall number of applications for the given years. Reassuringly, for the Protivínský and Korbel (2023) datasets, these ratios amount to 50.08% and 50.71% for the years 2017 and 2021 respectively, which can be plausibly viewed as comparable to the values of my own.

[15] A possible explanation of this fact could be that the dataset provided by Protivínský and Korbel (2023) was also filtered according to this criterion; the file outlining the cleaning process submitted to the author, however, does not mention this. In any case, I would expect the proportion of filtered observations to be comparable to the 2007 and 2011 datasets.

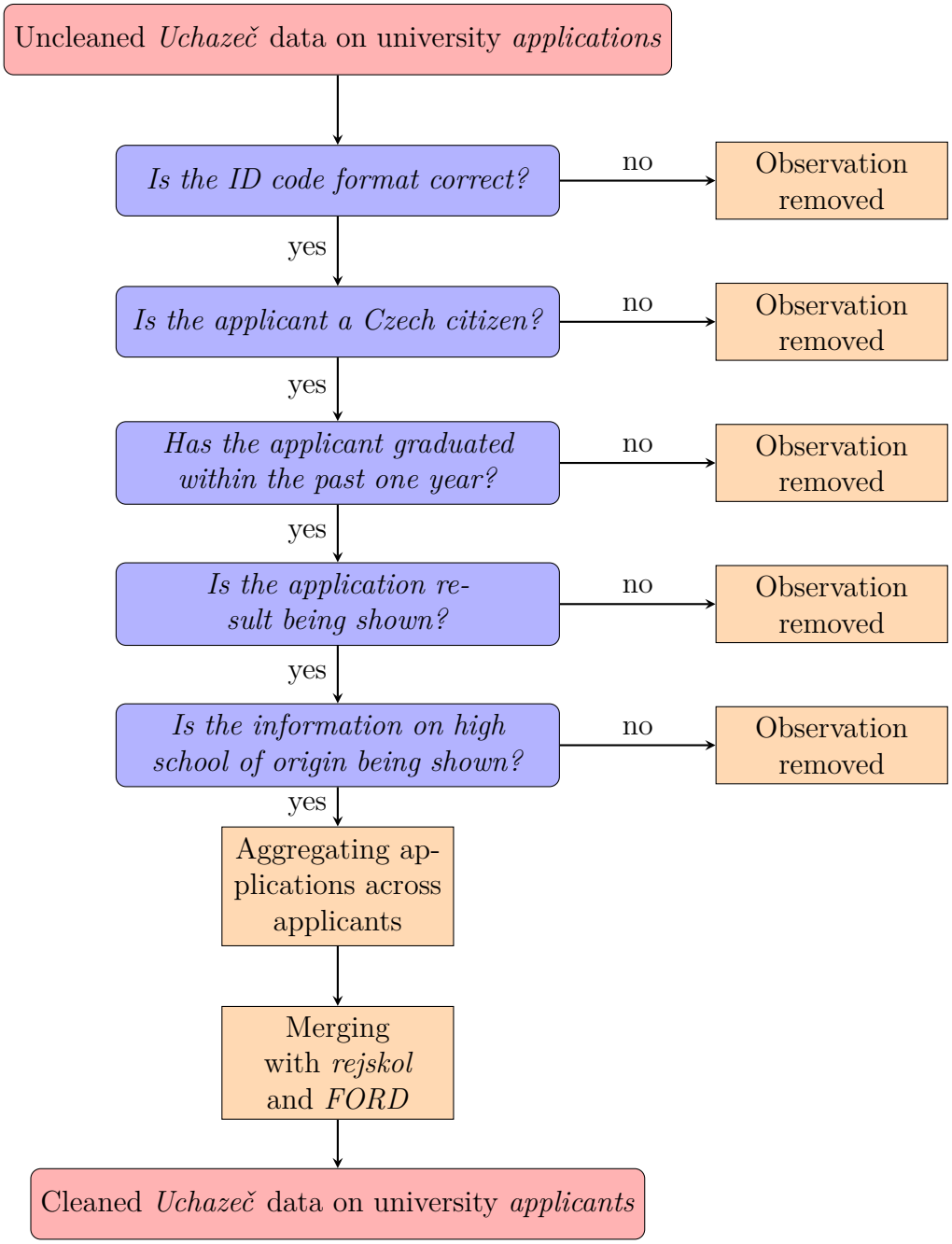

Figure 1: Cleansing Flowchart for the *Uchazeč* Database of University Applications

From Applications to Applicants

Having ensured that only the relevant and reliable observations are included in our dataset, the next step in the preparation of my data is to move from the level of *applications* to the level of *applicants*. As in many other countries, the number of tertiary degree applications an individual can submit is not limited in the Czech Republic. In the *Uchazeč* dataset, we can therefore potentially observe a multiplicity of row-observations for every applicant, each one of them corresponding to an individual application of the individual.

To account for this, a natural approach would be to simply use the applicant ID as a unique identifier, condensing the information from all of the individual's applications into a single row-observation. One property of the *Uchazeč* dataset, however, makes this relatively straightforward step a bit more tricky to perform. As I have already hinted at above, given the sensitive nature of the data, the applicant ID numbers have been anonymised in the source file. As a result, it is indeed possible that given the random process of ID anonymisation, two or more applicants could in theory share a single, anonymised ID. Aggregating across these, this could potential result in additional applications being incorrectly ascribed to a single applicant ID.

While this could plausibly be viewed as a random error, therefore not presenting a substantial threat to my subsequent quantitative inference, I chose to take the following steps to mitigate its occurrence. Assuming that the remaining individual-level variables, e.g., the high school of origin (*IZO* code), the postal code *PSČ*, or nationality (*stát* code), are stationary, grouping by such variables during the aggregation step would result in the individual applicant being uniquely identified. By conditioning on the *PSČ* variable, this is exactly the approach I take.

Admittedly, by means of the aggregation, changes in these variables *inter* individual applications could - in theory - result in that applicant being counted twice. Given the persistency of these variables, however, the likelihood of dou-

ble counting can plausibly be viewed as comparatively smaller relative to the 'anonymisation' random error.[16] In any case, considering the fact that even this error can be plausibly viewed as random, this arguably fails to present a notable issue.

III.B The STEM Attainment Measure

Very little has been written so far about a key component of the present research design: the identification of applications that can plausibly be classified as *STEM* and *non-STEM*. Unfortunately, the manner in which the term is being used in the STEM attainment literature is rather imprecise. Its usage in other research efforts can be very restrictive, delimiting the term to the areas of natural sciences and mathematics exclusively, or rather quite expansive, comprising the full breadth of medical sciences, computational economics, animal sciences, etc.

Classifying STEM Applications

An influential paper by Carrell et al. (2010),[17] for instance, uses "[m]ath and sciences" course outcomes in some portions of their analysis, which are purportedly aimed at assessing gender role-model effects in STEM. At the same time, though, Carrell et al. (2010) implicitly recognise that the exclusion of other fields - e.g., biology - is not self-evident. As Carrell et al. (2010, p. 1113) note, in the context of the United States, biological sciences "require less mathematics and have historically higher rates of female participation." It does not immediately follow, however, that biology should be excluded from the STEM category altogether.

The approach of Carrell et al. (2010) is to test for the professor gender effects using both definitions - an approach which is advantageous on the grounds of not being overly prescriptive, and of facilitating the cross-validation of the findings across substantively differential measures. With these virtuous considerations in

[16] It is, nonetheless, this specific consideration that led me to condition on only one of these 'stationary' variables, as each additional condition increases the likelihood of double counting. See the associated *R* code in Appendix II for details.

[17] See Section I.C for additional details.

mind, I choose to take this approach in this thesis also (see Section III.B below).

Given the lack of data availability for the STEM categorisation of individual programmes, however, the way in which I categorise each individual application as *STEM* or *non-STEM* relies on the focus of the *institution*, and not the *programme*. The reason motivating this decision is the following. While the *Uchazeč* dataset does provide an identifier of the specific study programme chosen by the applicant, no information is available on the broader area of study. In order to create an indicator specifying whether the field of study falls within the umbrella term of STEM or not, the researcher would have to manually categorise every single of the approximately 1,300 educational programmes according to their orientation of study.

Fields of Research and Development

Given this fact, I decided to pursue an alternative approach that would be viable within the scope of the research agenda of this thesis. Specifically, I opted to follow the methodology used in the production of national education, and research and development statistics by the ČSÚ (2023b), which also relies on an *institutional*, rather than *programme-based* field classification.

The approach exploited by ČSÚ (2023b) is as follows. For each research institution, the Czech Statistical Office tallies the number of researchers according to their main field of research. In accordance with the Frascati manual for the measurement of scientific, technological and innovation activities (OECD, 2015), there are six categories under which the researcher can be subsumed.[18] The ČSÚ then determines the main field of research[19] of the given institution by matching

[18]These are generally referred to as the *Fields of Research and Development* (FORD). To see their full list, and the STEM measure and ISCED-F conversion that are being used throughout this thesis, please refer to tables 1 and 2.

[19]Or, "*převažující vědní oblasti pracoviště*" in Czech (M. Štampach, personal communication, 17 July 2023, emphasis added).

it to the FORD categorisation of the plurality of its researchers.[20][21]

In order to ensure the reliability and comparability of my results to that of the ČSÚ and other authors, my original aim was to construct the application STEM measure using the FORD categorisation of universities and colleges as employed by the ČSÚ. Unfortunately, it was not possible to obtain the FORD classification due to the fact that according to the ČSÚ, the classification constitutes a confidential statistical information and cannot be provided to external researchers as such (M. Štampach, personal communication, July 17 2023).

In an effort to compensate for this, I attempted to mimick the ČSÚ methodology, exploiting a FORD categorisation measure of my own which I previously created over the course of a research assistantship project with prof. Ing. Štěpán Jurajda, Ph.D. While this categorisation measure was constructed to match the main field of research of the plurality of *PhD students* rather than *researchers* - a fact in part motivated by the availability of MŠMT performance data (MŠMT, 2023b) - these measures are likely to exhibit a significant overlap. Therefore, I would not expect the results to differ substantivelly under the ČSÚ classification.

One more note regarding this approach should be made, though. As the original (MŠMT, 2023b) classification is based on the *International Standard Classification of Education - Fields of Education and Training* (ISCED-F) categorisation (ČSÚ, 2013), an additional conversion had to be made between the ISCED-F and the FORD classification. While I deem this step to be fairly non-controversial, I nonetheless report the conversion key in table 1 for the reader's reference. In any case, it is hardly possible to think of an alternative conversion for the FORD 01 and 02 categories, which arguably matter the most for the construction of my preferred STEM*n* measure (see below). Additional information on this step, and

[20] I.e., if the given institution reports 60 researchers to be classified as *FORD 06 - Humanities and the arts*, and 40 classified as *FORD 05 - Social sciences*, the resultant classification of the institution's main field of research is *FORD 06 - Humanities and the arts*.

[21] Unfortunately, this ČSÚ methodology is not publicly accessible, and was provided to me through a personal correspondence with the Department of Research, Development, and Society of Information Statistics, or *Oddělení statistik výzkumu, vývoje a informační společnosti* in Czech (M. Štampach, personal communication, July 17 2023). The correspondence can be provided by the author upon request.

the detailed breakdown of the categorisation of universities and their components, are provided in Appendix I.

FORD Field	FORD	ISCED-F Field	ISCED-F
Natural Sciences	01	Natural sciences, mathematics and statistics	3
Engineering and technology	02	Information and communication technologies	6
		Engineering, manufacturing and construction	7
Medical and health sciences	03	Health and welfare	9
Agricultural and veterinary sciences	04	Agriculture, forestry, fisheries and veterinary	8
Social sciences	05	Education	1
		Social sciences, journalism and information	3
		Business, administration and law	4
Humanities and the arts	06	Arts and humanities	2

Source: ČSÚ (2013), OECD (2015), and author's own conversion.

Table 1: ISCED-F to Fields of Research and Development Conversion Table

STEM*n* and STEM*w* Measures

Finally, one additional step needs to be taken before the final, school-by-school STEM attainment measure can be computed: the creation of the requisite STEM dummy variable. Given the FORD categorisation methodology outlined above, I chose the natural option of subsuming the FORD categories which broadly corresponded to the definition of STEM under the umbrella of a binary variable. Considering the ambiguity in the usage of the term, however, I follow Carrell et al. (2010) in working with two particular conceptualisations of STEM fields: a *narrow* measure (or, STEMn) comprising the *Natural sciences*, and the *Engineering and technology* FORD fields, and a *wide* measure (or, STEMw) encompassing *Medical and health sciences*, and *Agricultural and veterinary sciences* in addition to the ones included in the STEMn definition. The conversion process is

illustrated in table 2, which can be found below.

FORD Field	**FORD**	**STEM*w***	**STEM*n***
Natural sciences	01	1	1
Engineering and technology	02		
Medical and health sciences	03		0
Agricultural and veterinary sciences	04		
Social sciences	05	0	
Humanities and the arts	06		

Source: OECD (2015), and author's own conversion.

Table 2: Fields of Research and Development to STEM Conversion Table

Considering the fact that the gendered differences appear to be especially pronounced in the fields of mathematics, physics, engineering and computer sciences (see the discussion in Section I), the STEMn, rather than the STEMw measure appears to be of primary importance to the relationships studied in this thesis. With that being said though, and similarly to what Carrell et al. (2010) remarks in the context of the United States, given the fact that the Czech Republic, too, has a history of women being comparatively more engaged in medical and biological professions, drawing the distinction between the STEMn and STEMw measures appears appealing. The advantage of doing so rests not only in increasing the robustness of the estimates for my preferred STEMn measure, but comparing the modelling outcomes for the STEMn and STEMw variables arguably additional nuance to the interpretation of the results.

From Applicants to High Schools

The aggregation across schools is, then, fairly straightforward. Treating each individual applicant as having applied to STEMn as long they submitted at least one application that was classified as STEMn,[22] I aggregate the number of all applicants, all female applicants, all STEM applicants, and all female STEM applicants across all high schools - a process made possible by the fact that the

[22]The aggregation process for STEMw is perfectly analogical.

MŠMT school registry *IZO* code constitutes a unique identifier of the given high school.

The resulting *schools* dataset, then, presents a baseline for the computation of the desired (female) STEM application ratio for each high school, defined as the number of (female) STEM applicants divided by the total number of (female) applicants, both in year t and high school i:[23]

$$\text{STEM application ratio}_{it} = \frac{\text{\# of all STEM applicants}_{it}}{\text{\# of all applicants}_{it}}$$

$$\text{Female STEM application ratio}_{it} = \frac{\text{\# of female STEM applicants}_{it}}{\text{\# of female applicants}_{it}}$$

All the while precluding the necessity to obtain additional data on class composition or school size, changes in the dependent variables thus conceptualised will allow us to assess the influence of socioeconomic and regional factors, extrapolate the time trends across all observations and - importantly - estimate the effects of policy measures, such as the SÚ AV ČR mentoring programme considered in this thesis. The specifics of this treatment variable, which is of particular interest vis-á-vis my primary hypothesis, are discussed in the following subsection.

III.C The Treatment

The treatment data originate from the Institute of Sociology at Czech Academy of Sciences (SÚ AV ČR) mentoring programme. The primary aim of the scheme was to make scientific work more attractive for women, and to support equal opportunities in research and development among high school students (NKC - gender a věda, personal communication, February 17, 2023). Building on experience from abroad, the conception of the mentoring was that of a one-to-one

[23]Note that this definition is applicable to both the STEMn and the STEMw measures, the only difference being in the number of students classified as STEM applicants in the numerator according to either measure.

mentoring, capturing the fact that one senior high school student - a mentee - is being guided by one university student - a mentor - majoring in a field with a history of female underrepresentation (Ibid.). The role of the mentor, then, was to introduce their mentee to the field in the context of tertiary education, and to provide advice and support to the junior student via regular meetings, thus assisting them in their academic and personal development.

According to the programme documentation, among the primary aims of the scheme were the goals of supporting girls with an interest in technical study fields, and - crucially - of questioning stereotypical gendered perceptions regarding women's study tracks (Ibid.). Remarkably, these objectives are in near-perfect alignment with the hypothesised mechanisms of a pro-STEM intervention as specified in Section II.A, and the mechanisms identified in the scholarly literature (see Section I).

As per the scope of the mentoring scheme, the programme ran from the 2010/2011 to the 2016/2017 academic year, thus spanning the full set of seven calendar years.[24] During this period, as much as 90 schools participated in at least one year of the programme. The data on participating schools were provided to the author via an evaluation report of the mentoring programme. Therefore, to operationalise the data for the purposes of estimation these needed to be coded and linked to the school register manually by the author. Additional details on this process can be found in the associated *R* code, as attached in Appendix II.

An overview of the number of schools that were exposed to the programme by region is, then, given in table 3 below. Reassuringly for the purposes of estimation, at least one school was exposed to the mentoring treatment in every region, although in some areas (e.g., *Pardubický kraj* or *Kraj Vysočina*), the exposure was comparatively limited.

As per the construction of the main treatment variable, this - too - can plausibly be viewed as fairly uncontroversial. Given that the treatment was limited

[24] Although admittedly, for the purposes of our main treatment variable, only observations for the years 2010/2011 and 2016/2017 coincide with the available *Uchazeč* datasets.

Region (*kraj*)	**No. of Participating Schools**
Hlavní město Praha	17
Středočeský kraj	11
Jihočeský kraj	2
Plzeňský kraj	4
Karlovarský kraj	2
Ústecký kraj	5
Liberecký kraj	3
Královéhradecký kraj	3
Pardubický kraj	1
Kraj Vysočina	2
Jihomoravský kraj	17
Olomoucký kraj	10
Zlínský kraj	5
Moravskoslezský kraj	8

Table 3: High Schools With an Exposure to the SÚ AV ČR Mentoring by Region

to one academic year, generally restricted to the penultimate or the ultimate year of high school study, I label a given school as treated for as long as at least one mentee from the given school participated in the programme in that year. Admittedly, considering the fact that penultimate year students were included in the treatment, introducing a lag in the treatment variable would be advisable; considering the fact that only an incomplete panel of the *Uchazeč* dataset was exploited in this thesis, it was unfortunately not possible to conduct this exercise.

To mitigate this concern, and to account for the possibility of medium-term to long-term effects of the treatment, I nonetheless conduct a robustness check using an *Ever Treated* (or, *evert*) variable. In specifications that employ this variable, I denote an observation as treated in period t for as long as it was treated in period t, or in any of the previous periods.

Before we proceed to the discussion of the final data source, an additional note on the treatment is in order. Admittedly, one could maintain that a scheme of this sort - i.e., one that is delivered via one-on-one consultations rather than a group intervention - is unlikely to have a considerable impact beyond that on the participating individual. While this contention is understandable, recalling my diccussion of the contemporary scholarly findings in Section I.C, there are

notable reasons to think otherwise.

As the literature suggests, peer effects can play an important role in questioning existing stereotypes and, as a consequence, affect female STEM engagement rates. Moreover, considering the role of institutional trust as documented by Moss-Racusin et al. (2018), the shared knowledge of students that their high school participates in a STEM encouragement programme could also prove emancipatory to other female students, prompting them to consider a STEM study track of their own. For these reasons in particular, I consider it likely that even a targeted programme of this sort can exhibit significant positive spillovers, thus making it sensible such a measure in a quantitative setting.

Finally, given the fact that the evaluating documents of the mentoring programme - on the basis of which the treatment measures was digitised and coded - are not considered publicly available material, these do not constitute an integral parts of this thesis. Additional details, and the opportunity to preview the SÚ AV ČR documents, will nonetheless be provided by the author upon request.

III.D MOS Municipality Controls

The final data source exploited in this piece is that of the Czech Statistical Office's (ČSÚ) *Municipality and district statistics* (MOS) database (ČSÚ, 2023a). This data source comprises open, municipality-level data on various socioeconomic, financial and structural properties of each individual municipality in the Czech Republic, covering the years from 2000 to 2022. These data available from the ČSÚ website, or upon direct request from the Office.

The main motivation for the inclusion of these municipality-level variables was to further facilitate the matching regression specifications, as detailed in Section IV.C, to provide additional means for controlling of time-varying factors in my difference-in-differences specification, and - crucially - to allow for the assessment of my socioeconomic hypothesis (see Section II.B).

Building on this last observation, in order to test for the hypothesis specifically

aimed at assessing the difference in STEM attainment for the schools located in metropolitan areas and those on the periphery, I use the MOS data to construct a city dummy variable specifically aimed at testing this relationship. While I consider the conversion fairly straightforward, I nonetheless outline my conversion methodology for the MOS municipality dummy in table 4 below.

Type	**ČSÚ Code**	**City Dummy**
Statutory cities and Prague	1	1
Cities	2	
Towns (*městyse*)	3	0
Non-chartered cities[25]	4	
Non-chartered towns[26]	5	
Military municipalities	9	
Other municipalities	0	

Table 4: Municipality Classification Conversion Table

Generally speaking, though, the variables that I chose to include in my analysis can roughly be divided in four areas: structural determinants (e.g., number of residents, road and rail length in the municipality, area, etc.), economic determinants (e.g., number of registered businesses of 249 or more employees, municipality tax income, corporate taxes accrued, etc.), and social determinants (e.g., number of social counselling facilities, or the proportion of unemployed residents). The inclusion of such controls, then, captures the fact that temporal changes in physical access to tertiary education, in economic situation of the area of residence, or the social dynamics of the region can also bear influence on tertiary education study decision. Hence, I view their inclusion in some of my more developed model specifications as appropriate.

Reassuringly, the MOS data source exhibited almost perfect overlap between the location (*Místo*) variable in the school registry database, and the MOS *obec* variable, allowing for a virtually flawless merging of the two data sources. The only exception to this was the capital city, Praha, which was broken down into administrative districts in the school registry database. For the purposes of merging the two sources, I attended to this issue by combining them into one district. Here again, I invite the interested reader to consult the *R* code in Appendix II

for details.

One final potential obstacle to inference could be seen in the changes in territory definitions (or, the addition of new ones) over the studied period.[27] Fortunately for my purposes, only 30 of municipality districts were modified in the relevant period. More reassuringy still, only one of them (the district of *Černovice*) reports having a high school, which is virtually impossible to bear any influence.

Finally, as the *Uchazeč* database and the MOS database use different numbering systems for the *kraje* regions - which are of high importance to our descriptive analysis in Section V.A - I conclude this Data section by providing a key for the conversion between the two categorisations. This conversion table is to be found in table 5 below.

Region (*kraj*)	**Code**	**CZ-NUTS 3**
Hlavní město Praha	0	CZ010
Jihomoravský kraj	1	CZ064
Jihočeský kraj	2	CZ031
Karlovarský kraj	3	CZ041
Kraj Vysočina	4	CZ063
Královéhradecký kraj	5	CZ052
Liberecký kraj	6	CZ051
Moravskoslezský kraj	7	CZ080
Olomoucký kraj	8	CZ071
Pardubický kraj	9	CZ053
Plzeňský kraj	10	CZ032
Středočeský kraj	11	CZ020
Zlínský kraj	12	CZ072
Ústecký kraj	13	CZ042

Table 5: Regions Classification Conversion Table

[27] Admittedly, an additional minor difficulty rested in the fact that a handful of municipalities (or, 2.85% to be precise) were listed twice, possibly due to the fact that some parts of the given municipality belonged to a second administrative district (*okres*). Considering the fact that our controls are at the level of municipality, and not at the level of *okres*, eliminating the duplicit entries should have no effect on the subsequent analysis.

IV Empirical Framework

In this section, I give a brief overview of the empirical modelling techniques employed to assess the hypotheses outlined in Section II.A. I chose a tested, and fairly conventional approach of starting with a simple pooled model, thus examining the general associative patterns included in the data, and gradually increase complexity from there by fine-tuning the model, including regional and time fixed effects, and additional controls. This effort culminates in the formulation of my two preferred specification: the difference-in-differences, and the matching regression.

IV.A A Set of Pooled Regressions

Far from having the ambition of being interpreted as causal, I first begin by fitting a simple, pooled regression of the following form. This approach allows us to calculate a 'naive', unobsfuscated difference in group means (DIGM) estimator for the treated and the untreated. By fully disregarding the time and the panel dimensions of the data, this can serve as a baseline of the difference between the two groups irrespective of any other influence:

$$y_{it} = \alpha + \delta T_{it} + \epsilon_{it} \tag{1}$$

In this, and in the equations that follow, y_{it} simply corresponds to the outcome variable of interest - i.e., the female STEMn (or, alternatively, female STEMw) application rate in school i at time t. The variable T_{it}, then, represents our treatment - i.e., an indicator stating whether the school i participated in the SÚ AV ČR mentoring programme in year t. As usual, α and ϵ_{it} capture the constant, and the error term, respectively.

Having formalised the DIGM estimator, I follow by incorporating the time dimension into my modelling through the inclusion of the time fixed, effects, θ_t

in my pooled regression:

$$y_{it} = \alpha + \theta_t + \delta T_{it} + \epsilon_{it} \tag{2}$$

I further add additional complexity by including the regional (*kraj*) fixed effects, μ_r, to account for the heterogeneity between regions, thus filtering out the context-specific regional influences on my dependent variable:

$$y_{irt} = \alpha + \theta_t + \mu_r + \delta T_{it} + \epsilon_{irt} \tag{3}$$

Finally, I decide to include a set of MOS and school controls, allowing us to finally gain an informed and more nuanced insight regarding the studied relationships:

$$y_{irt} = \alpha + \theta_t + \mu_r + \delta T_{it} + \mathbf{x}_{it}\beta + \epsilon_{irt} \tag{4}$$

In this specification, the $\mathbf{x}_{it}$ term corresponds to a vector of municipality and school-level controls, comprising the school type, the total number of applicants, the city dummy, the number of residents in the municipality, the size/length of its area, rail, road and motorway facilities, the number of social counselling facilities, the total number of businesses, the number of businesses of above 249 employees, personal and corporate taxes accrued, the tax income of the municipality, and the proportion of unemployed residents.

Even for this model with such an expansive set of controlling variables, one should note that the estimated coefficients cannot be immediately interpreted as causal. Arguably though, they provide at least some preliminary insight regarding the main hypothesised relationship. What is more, given the inclusion of the city dummy, and the economic variables, this specification additionally allows us to credibly assess the validity of my supplementary socioeconomic hypothesis through the analysis of the associated city and economic variables coefficients.

IV.B Difference-in-Differences Specification

As I note in the previous subsection, none of the pooled estimates can really be interpreted as causal. As usual, a part of the reason why this is the case in my setting rests in the possible presence of the omitted variable bias. Recalling the discussion of school- and class-level determinants of STEM attainment gap in Section I.C, one could, for instance, plausibly maintain that the fact that I do not control for the gendered composition of the teaching staff prevents me from attributing any observed effect to the mentoring treatment *per se*. To dispel the objections of the similar sort, and to facilitate causal interpretation, I therefore decided to exploit a difference-in-difference estimator - a strategy made viable by the panel character of my dataset.

Following Blumenau (2020), I implement the fixed-effects estimator to obtain the desired difference-in-differences estimand:

$$y_{it} = \gamma_i + \theta_t + \delta T_{it} + \epsilon_{it} \tag{5}$$

Here, in contrast to my previous specifications, γ_i is used to denote the fixed effect term for a given school i.

Turning for a moment to the issue of causal inference inside the difference-in-differences estimation frameworks, a crucial identifying assumption of the diff-in-diff estimator is that of *parallel pre-treatment trends*. Unfortunately, given the incompleteness of the *Uchazeč* panel, and the staggered timing of the treatment at different high schools, validating this assumption in my setting is difficult. In figure 2, I nonetheless provide a visual inspection of both STEMn, and female STEMn attainment rates for the studied period. In doing so, I differentiate between the schools which are classified as *Ever Treated* by the end of the studied period, and those classified as *Never Treated*:

As is apparent from the inspection of figure 2, the female STEMn ratio appears to fall at a lower rate between 2007 and 2017, while the reverse is true for the 2017 to 2021 period. Given the fact that some schools undergone treatment as early

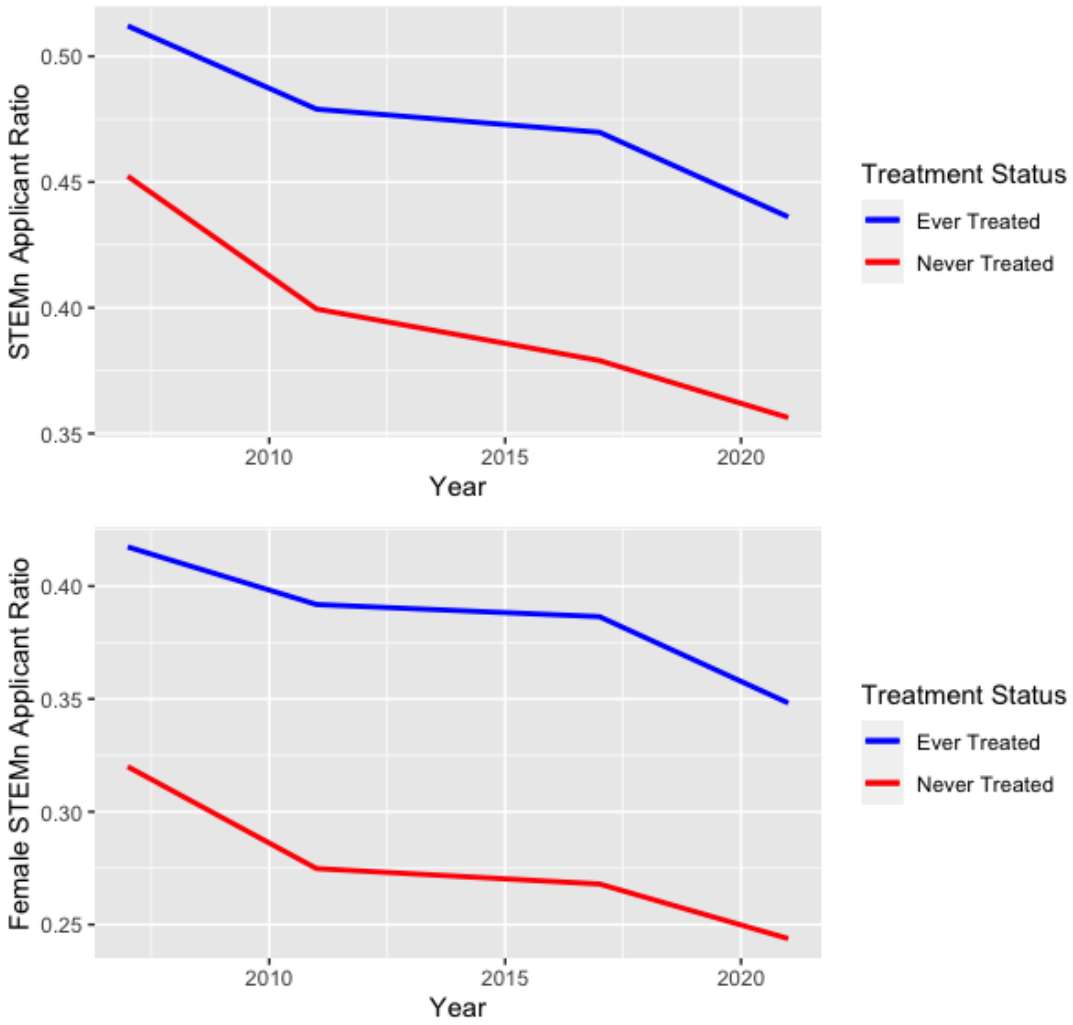

Figure 2: Line Chart of the STEMn Measure for the Ever Treated and the Never Treated (2021)

as in the school year 2010/2011, though, this does not immediately invalidate the usage of a difference-in-differences estimator. The reversed trend towards the end of the period, however, appears startling given the expected direction of the intervention's effect.

In addition, the notable differences between the overall *levels* of STEMn ratios for the two groups are, nonetheless, worrying. The employment of matching techniques, aimed to ensure that we are indeed 'comparing the comparables', therefore appears yet more appropriate.

IV.C A Matching Approach

The use of matching in this thesis can be motivated by two observations in particular: first, the striking imbalance between the number of observations for the *Never* and the *Ever Treated*, and the notable differences for other controls observables (including the differences STEMn levels overall, as discussed in the previous subsection)

An opportunity to inspect the underlying differences in more detail is presented to the reader in table 6 below. Not only does the number of controlling

	Never Treated	*Ever Treated*
No. of Observations	1,076	90
Tot. Applicants	38.72	83.21
Female Applicants	22.08	48.79
City Dummy	0.95	0.97
No. of Residents	271,603	303,006
Roads (km)	238.60	248.39
Rail (km)	203.39	217.37
No. of Businesses of >249 Employees	140.86	157.34
No. of Social Counselling Facilities	17.57	19.31
Personal Taxes Accrued ('000 CZK)	2,639,947	2,963,258
Corporate Taxes Accrued ('000 CZK)	3,236,983	3,633,578
Municipality Income ('000 CZK)	14,495,356	16,205,785
Proportion of Unemployed Residents (%)	2.48	2.55

Table 6: School-level Summary Statistics for the Ever Treated and the Never Treated Groups

observations outnumber that of the treated by a factor of 12, but as table 6 shows, we observe considerable differences in many other dimensions, too. Be it the gap in average total number of applicants, female applicants or municipality income, these factors can arguably be viewed as relevant to female STEMn attainment by virtue of affecting how the school operates, or by influencing the local dynamics.

As a complementary way to inspect the balance between the two sets of observations, I also offer a visual inspection of the distribution of the total number of applicants in figure 3, viewing it to be an especially relevant controlling variable by virtue of functioning as a proxy for the school size overall. As both figures in

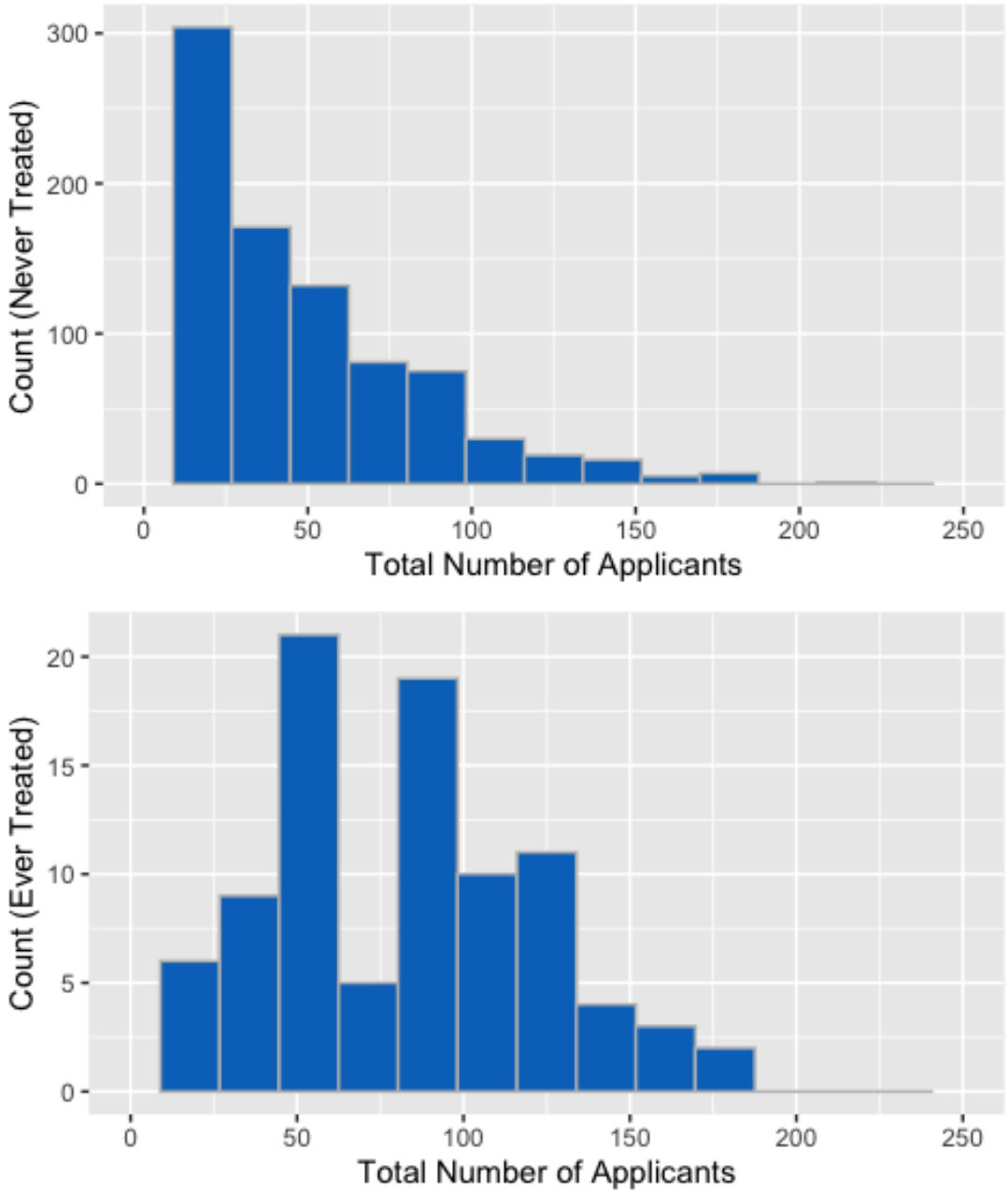

Figure 3: Histogram of 2021 Total Number of Applicants for the Never Treated and the Ever Treated

table 6 and the distributional visualisation of figure 3 document, the imbalance in observables between the two groups are remarkable. Given this inference, I decided to implement a matching design to ameliorate the concerns that in my previous specifications, we are 'comparing the comparables'.

Specifically, I implement a matching regression as proposed by Ho et al. (2011), operationalised via the associated *MatchIt* package in *R*. As I maintain,

through my chosen employment of a one-to-one, nearest distance GLM matching method, the difference in group means between the treated and the non-treated will better capture the true causal effect of being treated compared to my previous specifications.

As per my chosen matching variables, these for the most part correspond to the control variables exploited in the pooled regression models,[28] comprising the school type, the region (*kraj*) within which the school is located, the total number of applicants, the city dummy, the number of residents, the size/length of its area, rail, road and motorway facilities, the number of social counselling facilities, the total number of businesses, the number of businesses of above 249 employees, personal and corporate taxes accrued, the tax income of the municipality, and the proportion of unemployed residents.

Having concluded the formalisation of my main model specifications, I am now in a position to introduce the results of my quantitative modelling. That is the subject of the following section.

V Results

In this section, I present the results of my three main model specifications, as outlined in Section IV: the pooled, the difference-in-differences, and the matching model. For each specification, I present the results of my preferred, STEMn measure, and the more general STEMw. Before I examine the outcomes of my econometric estimation though, I briefly comment on the general trends and descriptive statistics of the female STEM application measures, and the STEM applicant ratio overall. These trends, I maintain, are worth paying attention to considering the fact that the *Uchazeč* dataset has not been previously studied in this manner.

[28] The motivation for their inclusion is - beyond the prevalent imbalances documented in table 6 - analogical.

V.A General Trends

In the recent past, it has become a stylised fact that the proportion of STEM applicants in Czech tertiary education has been steadily falling over the past two decades. As my findings document, this contention appears to be, at least in part, grounded in actual data. Capturing the proportion of applicants choosing to study a STEM study programme for the entirety of the Czech Republic, table 7 documents the falling trend in the STEMn attainment measure - a pattern that emerges while analysing both total and, to a lesser extent, female application ratios.

	2007	**2011**	**2017**	**2021**
STEMw/Total Applicants (%)	54.82	52.18	54.65	53.11
Female STEMw/Total Female Applicants (%)	44.77	43.70	47.83	46.89
STEMn/Total Applicants (%)	44.79	40.79	39.58	37.35
Female STEMn/Total Female Applicants (%)	32.18	29.08	29.15	26.70

Table 7: Development of the Proportion of STEM Applicants over Time

What is intriguing about these patterns, though, is that while the proportions of STEM applications *as measured by STEMn* steadily falls, the STEMw measure remains more or less stationary, or even rises slightly for female application rates over the whole studied period. This draws out an interesting distinction between the students' falling interest in 'hard' sciences such as mathematics, physics, or chemistry, and - possibly increasing - interest in health, medical and agricultural fields.

Concurrently, the breakdown of these trends across Czech regions (*kraje*) - the governments of which are responsible for the management of a majority of high-school institutions in the administrative area - also draws out a number of notable patterns.

As figure 4 demonstrates, in line with the general trends captured in table 7 we can observe the overall STEM application declining. What is remarkable, though, is that the reduction in STEM application rates is far from uniformly distributed across Czech administrative regions. While in *Kraj Vysočina*, the

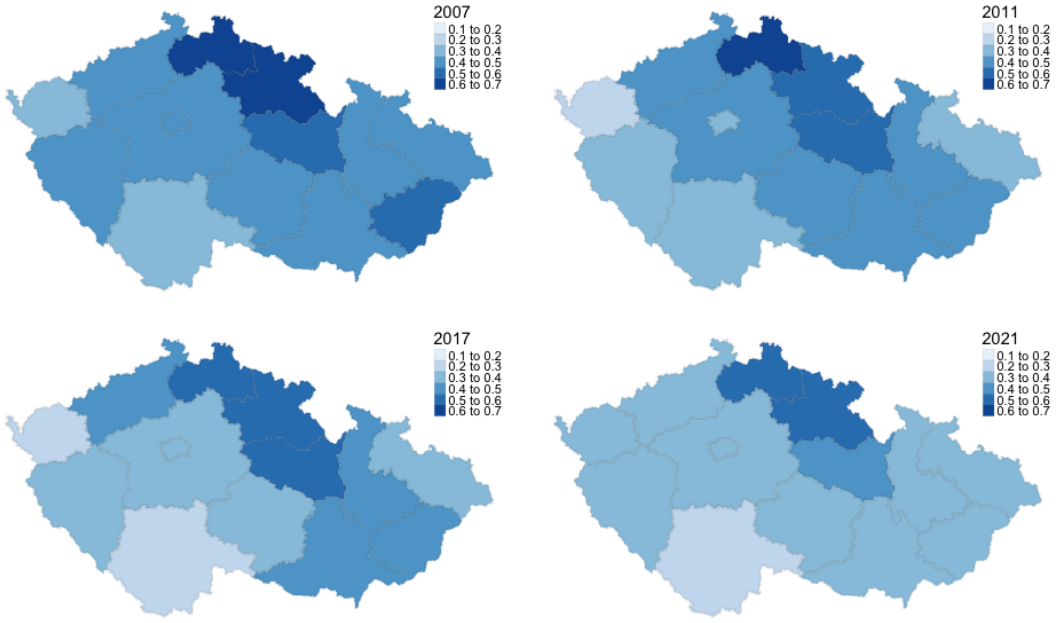

Figure 4: Proportion of STEMn Applicants across Regions and Time

proportion of STEMn applicants fell by as much as 16.4pp over the studied period, in *Karlovarský kraj* the ratio remained more or less constant, experiencing a fall of merely 0.4pp between 2007 and 2021.

The regional heterogeneity in temporal STEMn attainment are further reinforced through the analysis of *female* STEMn applicant rates. While for the *Olomoucký kraj*, we observe a fall in female STEMn measure of 6.9pp over the studied period, figure 5 documents that in *Karlovarský kraj*, the proportion of female STENn applicants actually *rose* by the margin of 7.5pp. Given the scope for influence on secondary education outcomes on the part the regional governments, such striking differences may be worth exploring further in some future work.

Two final notes are in order here. First, while for the sake of convenience I choose to report the visuals for my preferred STEMn measure, the STEMw can, too, provide additional information and nuance to the observed STEM applicant trends. I provide these visuals in Appendix II for the reader's reference.

Second, for both figure 4 and figure 5, I compute the region- and year-specific STEMn proportion by taking the total number of (female) STEMn applicants,

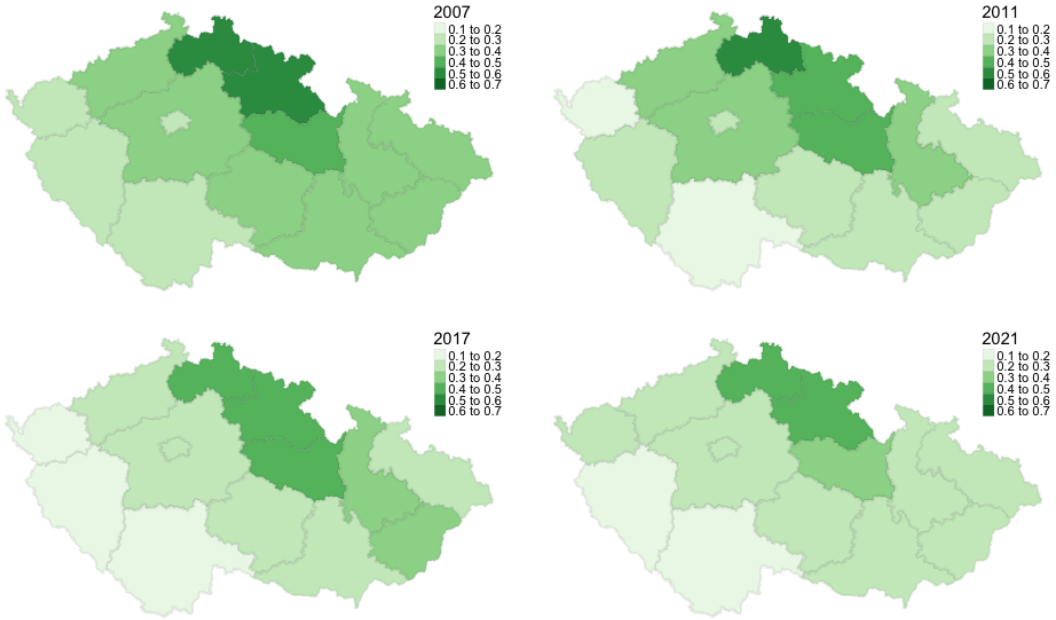

Figure 5: Proportion of Female STEMn to Female Applicants across Regions and Time

dividing it with the total number of (female) applicants. Alternatively, one could also take the average across *school* proportions for each region, rather than than computing the proportion for each region-year observation individually. This is not pursued here, although the 'average across schools' measure could conceivably be of more interest to other researchers.

V.B Main Findings

The findings from my main modelling specifications are shown below. In analogy to Section IV, I begin with an overview of the pooled regression results for both the STEMn and the STEMw measures, which are then followed with an overview of the difference-in-differences estimates. Finally, this subsection concludes with the display of my 2011 and 2017 matching results for the *Treated* group.

Pooled Regressions

Table 8 reviews the results from the set of pooled regressions, implementing the models specified in equations 1 to 4 respectively. As can be seen, while all of the

treatment coefficient estimates are positive - a result in line with the hypothesised effect of the intervention - none of them are statistically significant for any conventional α level.

	Dependent variable:			
	STEMn			
	(1)	(2)	(3)	(4)
Treated	0.030	0.029	0.023	0.008
	(0.049)	(0.049)	(0.049)	(0.050)
Total No. of Applicants				0.001***
				(0.0002)
City Dummy				0.062*
				(0.032)
Time Fixed effects	No	Yes	Yes	Yes
Region Fixed effects	No	No	Yes	Yes
Municipality Controls	No	No	No	Yes
Observations	2,861	2,861	2,807	2,805
R^2	0.0001	0.0002	0.012	0.027
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: STEMn Pooled Regression Results

Notably though, in the fourth of my pooled specifications, the *Total No. of Applicants* and the *City Dummy* controls are both significant at $\alpha = 0.05$ and $\alpha = 0.01$ level respectively, giving a preliminary assessment of my metropolitan hypothesis.

Replication of these results for the STEMw dependent variable, on the other hand, further confirms the inference of the STEMn models. Here again, as table 9 confirms, the signs of the treatment coefficients are correct, although - as in previous instance - statistically insignificant. What is more important, though, the *City Dummy* coefficient becomes significant at the $\alpha = 0.01$ level, lending further credence to the metropolitan hypothesis of section II.B.

	Dependent variable:			
	STEMw			
	(1)	(2)	(3)	(4)
Treated	0.041	0.037	0.036	0.032
	(0.048)	(0.048)	(0.049)	(0.049)
Total No. of Applicants				0.0004**
				(0.0002)
City Dummy				0.090***
				(0.030)
Time Fixed effects	No	Yes	Yes	Yes
Region Fixed effects	No	No	Yes	Yes
Municipality Controls	No	No	No	Yes
Observations	3,053	3,053	2,983	2,981
R^2	0.0002	0.001	0.006	0.026
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: STEMw Pooled Regression Results

Difference-in-Differences

Turning now to the difference-in-differences specification, the results for both the STEMn and the STEMw measure are being shown in table 10 below.

Unfortunately, the treatment coefficient estimate turns negative for the STEMn specification, although reassuringly, it remains highly statistically insignificant. Similarly, the STEMw model also fails to yield statistically significant results, although the sign of the coefficient is in this instance in alignment with the intervention hypothesis. Admittedly, based on these results alone, it would not be possible to reject the null hypothesis for the main hypothesised relationship between the intervention and increased female STEM application rates.

	Dependent variable:	
	STEMn	STEMw
	(1)	(2)
Treated	−0.006	0.015
	(0.038)	(0.034)
2011	−0.014	0.002
	(0.009)	(0.008)
2017	−0.007	0.004
	(0.010)	(0.008)
School Fixed Effects	Yes	Yes
Observations	2,861	3,053
R^2	0.764	0.803
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 10: Difference-in-Differences Regression Results

Matching

Examining the final set of my results, I turn to the matching regressions, as specified in section IV of this thesis. Given the incompleteness of the *Uchazeč* panel and the timing of the treatment, it was possible to implement matching only for the 2011 and the 2017 cross-sections for the *Treated* variable. The results of these models are reported in tables 11 and 12 respectively.

Studying the 2011 cross-section first, as documented in table 11, the sign for either STEMn or STEMw measure is positive, which is encouraging. In this specification also, however, the coefficients are insignificant, preventing me from rejecting the null hypothesis for my main, intervention hypothesis. Given the fact that I employed a one-to-one matching technique, and that only eleven schools were treated in the 2011 period, the fact that the estimands for the treatment variable are insignificant in this setting is not that surprising after all.

Interestingly though, the 2017 cross-section reveals a number of significant findings. As captured by table 12, not only are the treatment coefficients for both STEMn and STEMw positive and large in magnitude, but they are also

	Dependent variable:	
	STEMn	STEMw
	(1)	(2)
Constant	0.448***	0.494***
	(0.076)	(0.079)
Treatment	0.052	0.041
	(0.108)	(0.112)
Observations	22	22
R^2	0.012	0.007
Adjusted R^2	−0.038	−0.043
Residual Std. Error (df = 20)	0.253	0.262
F Statistic (df = 1; 20)	0.237	0.138
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 11: Matching Regression Results for the 2011 Mentoring Treatment

both statistically significant. These encouraging results allow me to reject the null hypothesis of no relationship between intervention and the STEMn and STEMw measures at the level of $\alpha = 0.1$ and $\alpha = 0.05$ respectively in this particular setting.[29]

Naturally, one needs to view these results in the context of the results obtained in the remaining regression specifications. The fact that a commonly employed method of causal inference shows signs of a statistically significant relationship represents a finding that should not be viewed as entirely meaningless. This is especially so given the robustness discussion that directly follows.

V.C Robustness Checks

First, one should note that the fact that in all of my specification I am already including both the STEMw and the STEMn measure should be viewed as a quasi-robustness check in its own right. Arguably, the inclusion of both measures allows me to see whether the results are driven mostly by the way in which the

[29] Although admittedly, the standard level of significance that is conventionally accepted in economic literature is that of $\alpha = 0.05$.

	Dependent variable:	
	STEMn	STEMw
	(1)	(2)
Constant	0.400***	0.432***
	(0.037)	(0.036)
Treatment	0.101*	0.133**
	(0.052)	(0.051)
Observations	66	66
R^2	0.056	0.095
Adjusted R^2	0.041	0.081
Residual Std. Error (df = 64)	0.211	0.208
F Statistic (df = 1; 64)	3.805*	6.755**
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 12: Matching Regression Results for the 2017 Mentoring Treatment

dependent variable was defined, or whether they can indeed be attributed to some fundamental property of the studied data. Given the fact that for the most part, the STEMn and STEMw results are in agreement with one another, this appears to lend credence to the contention that the estimates, while largely insignificant, are truly data driven.

Second, to check for the longevity of the treatment effect and to see whether the results are indeed time-contingent rather than spurious, I re-estimate my main models using the *Ever Treated* variable (see section III.C for the definition). If it were the case that the treatment is very time specific or short-lived, then we would expect the magnitude of the estimated relationship to diminish. The motivation for this inference is fairly straightforward: as the units that are 'authentically' treated in period t are mixed together with the ones that were treated in one of the previous periods, the strength of the relationship weakens.

Indeed, as table 13 shows, the *Ever Treated* effect is considerably smaller in magnitude relative to the *Treated* in table 8. In a similar vein, in the difference-in-differences specification, we observe the STEMw effect turning negative, whereas STEMn becomes increasingly more so - an effect consistent with the 'dilution' of

the treatment. These results, and additional details, are shown in table 14,

	Dependent variable:			
	STEMn			
	(1)	(2)	(3)	(4)
Ever Treated	0.007	0.007	0.009	−0.005
	(0.024)	(0.024)	(0.024)	(0.025)
Total No. of Applicants				0.001***
				(0.0001)
City Dummy				0.068**
				(0.027)
Time Fixed effects	No	Yes	Yes	Yes
Region Fixed effects	No	No	Yes	Yes
Municipality Controls	No	No	No	Yes
Observations	3,844	3,844	3,756	3,752
R^2	0.00002	0.0001	0.012	0.027
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13: STEMn Pooled Regression Results for the Ever Treated

Reassuringly though, the 2017 matching specification coefficients remain positive and statistically significant, suggesting that in 2017 - being a year in which the mentoring programme culminated, spanning the full breadth of 33 participating high schools - the treatment (or its correlates) could have been indeed successful in stirring female STEM engagement.[30][31]

Finally, one issue that was not dealt with during the data cleansing process (see section III.A for details) was that of missingness in additional information on high school of origin. This could potentially be an issue as the high school missingness in *location* is quite substantial in some years. For instance, in the 2007

[30]Recalling the discussion in section III.C, an alternative explanation for the significance of the treatment effect could also rest in the fact that the treatment may exhibit a one-year lag. If this were the case, the results could be driven by the 2016, rather than the 2017 treatment. This is plausible, as a comparable number of cohorts were treated in 2016, and some of which may enter their final year in 2017.

[31]As a side note, be advised that the 2011 matching regressions results for the *Ever Treated* are not being displayed due to the fact that the 2011 *Ever Treated* observations fully coincided with the 2011 *Treated* observations.

	Dependent variable:	
	STEMn	STEMw
	(1)	(2)
Ever Treated	−0.037	−0.011
	(0.023)	(0.021)
2011	−0.014	0.001
	(0.010)	(0.008)
2017	−0.003	0.007
	(0.010)	(0.009)
2021	−0.014	0.003
	(0.010)	(0.009)
School Fixed Effects	Yes	Yes
Observations	3,844	4,123
R^2	0.713	0.753
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 14: Difference-in-Differences Regression Results for the Ever Treated

	Dependent variable:	
	STEMn	STEMw
	(1)	(2)
Constant	0.401***	0.459***
	(0.037)	(0.036)
Ever Treated	0.090*	0.095**
	(0.048)	(0.047)
Observations	66	66
R^2	0.053	0.061
Adjusted R^2	0.038	0.046
Residual Std. Error (df = 64)	0.189	0.186
F Statistic (df = 1; 64)	3.577*	4.154**
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 15: Matching Regression Results for the 2017 Ever Treated

and the 2011 period, as much as 49.36% and 28.52% of observations respectively did not show an information on their location.[32]

While a non-negligible portion of such observations are eliminated throug the filtering process and missingness should not therefore invalidate - at least in theory - my main difference-in-differences specification, it does present an obstacle to other approaches to modelling. Therefore, to assess whether the inclusion of such observations significantly affects my results, I reestimate my main models excluding the school observations that do not include an information on their location. Reassuringly, the results in all main specifications are virtually identical, suggesting that the high school information missingness is unlikely to bias the substance of my estimates.[33]

VI Discussion

Over the course of the previous section, I have presented to the reader the quantitative estimates of the main empirical models developed in this thesis. In this penultimate section, it is my aim to briefly restate my main findings, and - crucially - to offer a key to the interpretation of my, possibly difficult to synthesise estimates. Subsequently, this section follows with a short discussion of the overall limitations of my empirical strategy and the results, concluding with the contributions of this work to the wider labour economics discipline.

As section V revealed, while the direction of the vast majority of my treatment estimates were in alignment with the intervention hypothesis, the set of my statistically significant results remains limited. Admittedly, the notable exception to this fact is the 2017 matching regression, which revealed a statistically significant relationship between the intervention and the STEMw measure, postulating that

[32] Admittedly though, the more recent datasets appear to exhibit much lowever levels of location missingness - for 2017 and 2021, "only" 19.40% and 13.32% observations respectively failed to report high school location.

[33] The associated R code employed in the performance of this check can be found in Appendix II; the detailed results of this robustness check, however, do not consist an integral part of this thesis, and are available upon request from the author.

being assigned to treatment in 2017 was associated with a 13.3*pp* average increase in the proportion of female students applying for a STEM*w* subject.

Yet concurrently, the difference-in-difference estimator found no statistically significant relationship overall. How to reconcile these contradictory findings? Arguably, a part of the solution that could have motivated the unsatisfactory results of the difference-in-differences design may rest in the properties of the *Uchazeč* dataset, and the limited scope of the SÚ AV ČR tretament. I discuss these in the Limitations subsection that follows. At this point, however, it suffices to say that given the contradiction in the two model specifications, it is not possible to rule decisively on the validity of the intervention hypothesis. With that being said though, the results from my matching specification arguably do indicate that under specific conditions, an intervention can be associated with heightened female STEM application rates in a way that is statistically significant (see section VI.A for additional discussion).

So far, I have also said little about the validity of my supplementary socioeconomic hypothesis. Unfortunately, the only one of my empirical designs that was suitable for its assessment was the pooled regression design, which arguably presents an obstacle to the causal interpretation of my estimates. With that being said though, the fact that the *City Dummy* variable has proved to be significant in both the STEM*n* and the STEM*w* specifications (including the robustness exercises conducted in section V.C) does allow us to reject the null hypothesis of no relationship between the two concepts. We can therefore infer that according to the *Uchazeč* data as reported in this thesis, there indeed exists a statistically significant relationship between the metropolitan municipalities, and increased female STEM engagement rates. Whether this relationship can be viewed as merely associative or causal, however, is a subject for future research.

VI.A Limitations

Despite these encouraging findings, one may nonetheless have a number of reservations regarding the research design advanced in this thesis. In this subsection, I strive to discuss the most relevant of them, offering my comments regarding the validity of such objections.

First, building on my observations in section V, one may feel skeptical about the prospect of employing a difference-in-differences estimator in a setting where the *Uchazeč* school panel comprises only four, unevenly spaced years in a fourteen-year timespan. This is indeed a legitimate reservation. Recalling my discussion of the potentially short lifespan of the treatment effect, in a setting where the treatment spans only the years from 2011 to 2017 and where the early treatment years are characterised by a low number of treated observations, this can potentially lead to the treatment effect becoming insignificant.

Unfortunately, within the scope of this research project which was severely restricted by the data availability of the *Uchazeč* data, there are very few alternatives to remedy this. Moreover, the potentially viable option of treating each individual cohort as a cross-section appears hardly applicable in this research context. While such an approach could be useful to the study of individual applicants, it is hardly transferable to the context of high-schools, whose policies are likely to persist across many time periods.

Similarly, given my inability to control for the presence of other STEM pro-engagement policies, one may wonder whether any observed effect is indeed causally attributable to the intervention *per se*. Admittedly, this presents a valid concern - if it were the case that other pro-STEM engagement policies, or other school-specific influences coincided with the treatment, its effect would not be identified. Since I have not been able to control for this fact, it would not make much sense to interpret even the few significant results of this thesis as causal - a limitation that I have attempted to recognise throughout the discussion. This is not to say, however, that the associative relationships uncovered in some of my

specifications cannot be viewed as valuable, or indeed valid.

A final reservation could be held regarding the data reliability itself. Considering the fact that the public access to the *Uchazeč* data source is restricted and no outside methodology is provided by the MŠMT about the data source, one may deem it questionable whether the quality of data input can be ensured at all in my research effort.

While, admittedly, quality assurance does appear to be a potential issue here, a vast array of precautionary steps can be taken to ensure that any input errors are orthogonal to the estimation strategy. By *ex ante* filtering of the entries which have been shown to be the most problematic, the researcher can ensure that only data of sufficient quality enters the estimation.[34] In the context of this thesis specifically, given the additional fact that excluding some of the problematic entries *ex post* does not affect my estimates in a systematic way either (see section V.C for details), one could plausibly maintain that even in the event that the data input *were* to an extent unreliable, the unreliability consists a random, and consequently harmful influence on the research project advanced in this thesis.

VI.B Contribution

Before I conclude, allow me briefly to review the main contributions of this work. The first, and arguably the most notable contribution of this thesis, consists in the novel use of previously unavailable data sources. Representing one of the first serious attempts to provide an in-depth analysis of the *Uchazeč* datasets - including the years of 2007 and 2011, which previously remained unstudied - the value of this thesis rests both in the extensive process of data preparation, and in the provision of a first-time preview of the patterns and trends in the behaviour of Czech tertiary education applicants.

Second, by linking the *Uchazeč* data with previously non-digitised SÚ AV ČR mentoring treatment data, I provide a quantitative evaluation of this specific

[34] For an extensive discussion of this process vis-á-vis the *Uchazeč* data, see section III.A.

programme, serving as a baseline for evaluation of this particular scheme, and of programmes of a similar kind. Furthermore, considering the fact that I uncover significant results in some of my specifications, studying the import of comparable mentoring programme to the individuals beyond those that are actively involved could also be viewed as a valuable future research effort.

Finally, this thesis arguably present a contribution to the STEM attainment gap literature in two respects. First, through my descriptive analysis of the *Uchazeč* dataset, I have been able to map the recent trends in female STEM application rates for the Czech Republic - an important piece of information from the perspective of the overall gender attainment gap nationally.

Second, despite the fact that it was not possible for me to credibly single out the effects of the SÚ AV ČR mentoring from the possible concurrence of other mechanisms, the uncovered associative relationships between certain schools, their selection into treatment, and the heightened levels of female STEM engagement invites for additional research efforts in this area. Studying in more detail what sets these schools apart from others, the way they are being governed, and what other policies may have been enacted by them that prompted the significance of the observed relationship appears to present an intriguing and worthy scientific endeavour.

Conclusion

The persistent gender attainment gap continues to present a challenge to academia and to the policymaker. Due to its implications for both economic and personal wellbeing of both men and women, the analysis of its fundamental drivers, and the policy tools aimed to disarm them, rightly remains an ongoing research effort. In the field of STEM, this appears to hold especially true as female students and researchers continue to suffer from severe under-representation relative to the male in such disciplines - a pattern that begins to emerge as early as in university application decisions, as this work demonstrates.

In this thesis I have sought to analyse the evolution of these patterns across Czech regions, time, and - crucially - the extent to which STEM attainment gap can be manipulated by policy intervention and socioeconomic characteristics of the region. Exploiting the novel *Uchazeč* dataset, which had previously remained largely unstudied, this thesis attempted to provide new insight and analysis on the topic by cleaning and operationalising the 2007, 2011, 2017 and 2021 *Uchazeč* datasets. By linking them to high-school registry data (MŠMT, 2023a), municipality-level statistics and geospatial files on Czech administrative districts (ČSÚ, 2023a; ČÚZK, 2023), and SÚ AV ČR treatment data, I created STEMn and STEMw school-level measures capturing the proportion of (female) applicants who chose to apply for a STEM study field in their further study.

My descriptive analysis shows that since 2007, the proportion of STEM applicants - as measured by my preferred STEMn measure - fell from 44.79% to 37.35% in 2021 - a reduction of nearly 7.44pp over the studied period. While the fall is not as pronounced for the female STEMn applicant ratio, which experienced a reduction of 'only' 5.48pp, the reduction in STEM engagement rates is far from equally spread across Czech regions. Focusing on total STEMn applicants ratio, while *Karlovarský kraj* experienced a fall of only 0.4pp during the studied period, the proportion of STEMn applicants from *Kraj Vysočina* fell by as much

as 16.4*pp* over the same timespan. Concurrently, an analysis of female STEM engagement rates reveals a very similar pattern.

As per the regression analysis, employing a combination of pooling, difference-in-differences, and matching models, I find a positive relationship for most specifications, although admittedly, the difference-in-differences estimates turn out negative for three of the four specifications. Unfortunately, virtually none of the coefficients proved to be statistically significant, the sole exception being the 2017 matching regression design. Notably, though, in this specification, I find that undergoing the SÚ mentoring treatment in 2017 was associated with a 10.1*pp* and 13.1*pp* increase in STEM engagement as measured by the respective STEMn and STEMw variables - a result that was significant at the 0.10 and 0.05 alpha level of significance, respectively. Note, however, that this result needs to be taken with a grain of salt, as none of the other specifications confirm this inference with a similar degree of statistical significance.

Turning to my analysis of the municipality-level covariates, my pooled regression results demonstrate a statistically significant difference between schools located in cities, relative to these in other areas, and at schools with larger pools of applicants. Both of these characteristics are positively associated with female STEM engagement, thus preventing me from rejecting my second hypothesis regarding the influence of socioeconomic characteristics on female STEM engagement. One should note, though, that the effect of applicant pool size is relatively small, and that the pooled specification prevents us from interpreting these results as causal due to the panel nature of the *Uchazeč* dataset and due to endogeneity concerns. In any case, the difference between metropolitan and non-metropolitan schools, and their associated dynamics, are arguably a topic worth exploring further in future research.

Despite the fact that my specifications have uncovered only limited statistically significant evidence in favour of a long-lasting, positive effect of the SÚ AV ČR mentoring programme treatment, this thesis has made a number of notable

contributions to the field of attainment gap research in the Czech Republic. Having operationalised, and subsequently employed the novel, previously unexploited *Uchazeč* data source, I have been able to validate previous, mostly circumstantial evidence on medium-term STEM engagement patterns and to add a regional dimension to the existing scholarly knowledge. Similarly, my efforts in linking it to data from other sources, including ČSÚ, ČZÚK, and MŠMT databases, presents a valuable addition in that it offers a new potential avenue of inquiry for future researchers wishing to make a contribution to the field of labour or education economics.

Finally, the fact that the findings do not fully invalidate my main hypothesis and can - to some extent - be attributable to the shortcomings of the treatment data and of the *Uchazeč* source, speaks in favour of analysing the impacts of pro-engagement policies on female STEM engagement in a more favourable setting. Indeed, the fact that even in this limited setting a credible estimation strategy such as matching uncovers statistically significant evidence suggests that further research efforts studying the influence of high-school intervention measures, or the school environment in general, appear to be a worthy endeavour in furthering our understanding of the attainment gap mechanics.

Bibliography

Anderson, D. J., Binder, M., & Krause, K. (2002). The motherhood wage penalty: Which mothers pay it and why? *American economic review*, *92*(2), 354–358.

Bedard, K., Dodd, J., & Lundberg, S. (2021). Can positive feedback encourage female and minority undergraduates into economics? *AEA Papers and Proceedings*, *111*, 128–132. https://doi.org/10.1257/pandp.20211025

Bell, L. A. (2005). Women-led firms and the gender gap in top executive jobs.

Bertrand, M., & Hallock, K. (2001). The gender gap in top corporate jobs. *Industrial & Labor Relations Review*, *55*(1), 3–21. https://doi.org/10.2307/2696183

Bishu, S. G., & Alkadry, M. G. (2017). A systematic review of the gender pay gap and factors that predict it. *Administration & Society*, *49*(1, SI), 65–104. https://doi.org/10.1177/0095399716636928

Blumenau, J. (2020). Panel data and difference-in-differences [Lecture notes.]. University College London.

Brenøea, A. A., & Lundberg, S. (2018). Gender gaps in the effects of childhood family environment: Do they persist into adulthood? *European Economic Review*, *109*(SI), 42–62. https://doi.org/10.1016/j.euroecorev.2017.04.004

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144. https://doi.org/10.1162/qjec.2010.125.3.1101

ČSÚ. (2013). Klasifikace oborů vzdělání (cz-isced-f 2013) [fields of education classification (cz-isced-f 2013)]. https://www.czso.cz/csu/czso/klasifikace-oboru-vzdelani-cz-isced-f-2013

ČSÚ. (2023a). Databáze MOS [MOS database]. https://www.czso.cz/csu/czso/databaze-mos-otevrena-data

ČSÚ. (2023b). Ukazatele výzkumu a vývoje - 2021 [Research and development indicators - 2021]. https://www.czso.cz/csu/czso/ukazatele-vyzkumu-a-vyvoje-2021

ČÚZK. (2023). Vybraná data rúian poskytovaná pro stát ve formátu shp [Selected RÚIAN state-level data in the shp format].

Dvořák, T., Zouhar, J., & Treib, O. (2022). Regional peripheralization as contextual source of populist attitudes in germany and czech republic. *Political Studies*, https://doi.org/10.1177/00323217221091981. https://doi.org/10.1177/00323217221091981

Evans, M. O. (1992). An estimate of race and gender role-model effects in teaching high-school. *Journal of Economic Education*, *23*(3), 209–217. https://doi.org/10.2307/1183223

Federičová, M. (2019). *How school report grades affect pupils' life decisions* (tech. rep.). IDEA at Centre for Economic Research and Graduate Education - Economic Insitute.

Fitzenberger, B., Schnabel, R., & Wunderlich, G. (2004). The gender gap in labor market participation and employment: A cohort analysis for West Germany. *Journal of Population Economics*, *17*(1), 83–116. https://doi.org/10.1007/s00148-003-0141-6

Gevrek, Z. E., Gevrek, D., & Neumeier, C. (2020). Explaining the gender gaps in mathematics achievement and attitudes: The role of societal gender equality. *Economics of Education Review*, *76*. https://doi.org/10.1016/j.econedurev.2020.101978

Hedija, V. (2016). Gender wage differences in the Czech public sector: A micro-level case. *Review of Economic Perspectives*, *16*(2), 121–134. https://doi.org/10.1515/revecp-2016-0009

Hegewisch, A., Liepmann, H., Hayes, J., & Hartmann, H. (2010). Separate and not equal? gender segregation in the labor market and the gender wage gap. *IWPR Briefing Paper*, *377*, 1–16.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Jurajda, S., & Paligorova, T. (2009). Czech female managers and their wages. *Labour Economics*, *16*(3), 342–351. https://doi.org/10.1016/j.labeco.2008.11.001

Kuhn, P., & Shen, K. (2013). Gender discrimination in job ads: Evidence from China. *Quarterly Journal of Economics*, *128*(1), 287–336. https://doi.org/10.1093/qje/qjs046

Lavy, V., & Megalokonomou, R. (2019). *Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study* (tech. rep.). National Bureau of Economic Research.

Legewie, J., & DiPrete, T. A. (2009). Family determinants of the changing gender gap in educational attainment: A comparison of the US and Germany. *Journal of Contextual Economics–Schmollers Jahrbuch*, *129*(2), 169–180.

Marianne, B. (2011). New perspectives on gender, In *Handbook of labor economics*. Elsevier.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *Review of Economic Studies*, *76*(4), 1431–1459. https://doi.org/10.1111/j.1467-937X.2009.00551.x

Moss-Racusin, C. A., Sanzari, C., Caluori, N., & Rabasco, H. (2018). Gender bias produces gender gaps in STEM engagement. *Sex Roles*, *79*(11-12), 651–670. https://doi.org/10.1007/s11199-018-0902-z

Mouganie, P., & Wang, Y. (2020). High-performing peers and female STEM choices in school. *Journal of Labor Economics*, *38*(3), 805–841. https://doi.org/10.1086/706052

MŠMT. (2023a). Rejstřík škol a školských zařízení [School and school facilities registry]. https://rejstriky.msmt.cz/rejskol/

MŠMT. (2023b). Statistika výkonových ukazatelů veřejných a soukromých vysokých škol ČR [Performance statistics of public and private colleges and universities]. https://statis.msmt.cz/statistikyvs/vykonyVS1.aspx

Murgaš, F., & Klobučník, M. (2016). Municipalities and regions as good places to live: Index of quality of life in the Czech Republic. *Applied Research in Quality of Life*, *11*, 553–570.

OECD. (2015). *Frascati manual 2015*. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264239012-en

Petrongolo, B., & Ronchi, M. (2020). Gender gaps and the structure of local labor markets. *Labour Economics*, *64*. https://doi.org/10.1016/j.labeco.2020.101819

Porter, C., & Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American economic journal. Applied economics*, *12*(3), 226–254.

Prokop, D., Korbel, V., Dvořák, T., Marková, L., Gardošíková, D., Grossmann, J., Krajčová, J., & Münich, D. (2021). *Nerovnosti ve vzdělávání jako zdroj neefektivity [Inequality in education as a source of inefficiency]* (tech. rep.). PAQ Research and IDEA.

Protivínský, T., & Korbel, V. (2023). *Idea: Uchazeč* [Unpublished manuscript.]. Unpublished manuscript.

Protivínskỳ, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, *59*, 141–149.

Reinking, A., & Martin, B. (2018). The gender gap in STEM fields: Theories, movements, and ideas to engage girls in STEM. *Journal of New Approaches in Educational Research*, *7*(2), 148–153. https://doi.org/10.7821/naer.2018.7.271

Robb, R., & Robb, A. (1999). Gender and the study of economics: The role of gender of the instructor. *Journal of Economic Education*, *30*(1), 3–19. https://doi.org/10.2307/1183028

Rousille, N. (2023). *The central role of the ask gap in gender pay inequality* [Unpublished manuscript.]. Unpublished manuscript.

Schwab, K., Samans, R., Zahidi, S., Leopold, T. A., Ratcheva, V., Hausmann, R., & Tyson, L. D. (2017). The global gender gap report 2017. World Economic Forum.

Smith, M. L., & Sokolova, V. (2022). Gender gaps in educational pathways in the Czech republic. *British Journal of Sociology of Education*, *43*(2), 296–313. https://doi.org/10.1080/01425692.2021.1971062

Verdugo-Castro, S., Garcia-Holgado, A., & Cruz Sanchez-Gomez, M. (2022). The gender gap in higher STEM studies: A systematic literature review. *Heliyon*, *8*(8). https://doi.org/10.1016/j.heliyon.2022.e10300

Vohlídalová, M. (2021). The segmentation of the academic labour market and gender, field, and institutional inequalities. *Social Inclusion*, *9*(3), 163–174. https://doi.org/10.17645/si.v9i3.4190

Zajickova, D., & Zajicek, M. (2021). Gender pay gap in the Czech republic - its evolution and main drivers. *Prague Economic Papers*, *30*(6), 675–723. https://doi.org/10.18267/j.pep.787

List of Figures

List of Tables

List of Abbreviations

AV ČR = Czech Academy of Sciences
ČSÚ = Czech Statistical Office
ČÚZK = Czech Office for Surveying, Mapping and Cadastre
FORD = Fields of Research and Development
ISCED-F - International Standard Classification of Education - Fields of Education and Training
MŠMT = Ministry of Education, Youth and Sports of the Czech Republic
OECD = Organisation for Economic Co-operation and Development
pp = Percentage Points
STEM = Science, Technology, Engineering, and Mathematics
STEMn = Narrow definition of STEM field (see Section III)
STEMw = Wide definition of STEM field (see Section III)
SÚ AV ČR = Institute of Sociology of Czech Academy of Sciences
VŠ = Universities and college

Appendix I

The following faculties were subsumed under the respective ISCED-F category based on the plurality of their PhD students belonging to the given study field in 2021, as captured by MŠMT performance statistics (MŠMT, 2023b). In cases where a significant portion of the faculty's PhD students belonged to another field, this is noted in brackets.

- ISCED-F 1 – Education
  - 11410 PedF UK [signif 3]
  - 11510 FTVS UK [signif 9]
  - 12410 Pedagogická fakulta JUČB
  - 13430 Pedagogická fakulta UJEP
  - 14410 Pedagogická fakulta MU
  - 15410 Pedagogická fakulta UPO
  - 17450 Pedagogická fakulta OP
  - 18440 Pedagogická fakulta UHK [signif 2]
  - 23420 Fakulta pedagogická ZUP
  - 28150 Fakulta humanitních studií UTBZ
  - 76000 Univerzita Jana Amose Komenského Praha, s.r.o.
- ISCED-F 2 – Arts and Humanities
  - 11210 Filosofická fakulta UK [signif 1 and 3]
  - 11260 Katolická teologická fakulta
  - 11270 Evangelická teologická fakulta [signif 9]
  - 11280 Husitská teologická fakulta [signif 1 and 9]
  - 12210 Filozofická fakulta JUČB [signif 1 and 3]
  - 12260 Teologická fakulta JUČB [signif 9]
  - 13530 FUD UJEP
  - 13410 FF UJEP [signif 3 and 1]
  - 14210 FF MU [signif 3 and 1]
  - 15210 Filozofická fakulta UPO [signif 1 and 3]
  - 15260 Cyrilometodějská teologická fakulta [signif 3]
  - 17250 Filozofická fakulta OU
  - 17500 Fakulta umění OU
  - 18460 Filozofická fakulta UHK
  - 19240 Filozoficko-přírodovědecká fakulta SUO
  - 23330 Filozofická fakulta ZUP [signif 3]

  - 23410 Fakulta designu a umění L. Sutnara ZUP
  - 24520 Fakulta umění a architektury TUL [signif 7]
  - 25110 Fakulta restaurování UP
  - 25210 Fakulta filozofická UP
  - 26420 Fakulta výtvarných umění VUT
  - 28130 Fakulta multimediálních komunikací UTBZ
  - 51110 Hudební a taneční fakulta AMU
  - 51210 Divadelní fakulta AMU
  - 51310 Filmová a televizní fakulta AMU
  - 52000 Akademie výtvarných umění v Praze
  - 53000 Vysoká škola umělecko-průmyslová v Praze
  - 54510 Hudební fakulta JAMU
  - 54530 Divadelní fakulta JAMU
- ISCED-F 3 – Social sciences, journalism and information
  - 11230 Fakulta sociálních věd UK
  - 11240 Fakulta humanitních studií UK [signif 2]
  - 13510 Fakulta sociálně ekonomická UJEP [signif 4 and 9]
  - 14230 Fakulta sociálních studií MU
  - 14560 Ekonomicko-správní fakulta MU [signif 04]
  - 31150 Národohospodářská fakulta VŠE
  - 75000 Metropolitní univerzita Praha, o.p.s.
- ISCED-F 4 – Business, administration and law
  - 11220 PF UK
  - 12510 Ekonomická fakulta JUČB
  - 14220 Právnická fakulta MU
  - 15220 Právnická fakulta UPO
  - 19520 Obchodně podnikatelská fakulta v Karviné
  - 23320 Fakulta právnická ZUP
  - 23510 Fakulta ekonomická ZUP
  - 24310 Ekonomická fakulta TUL
  - 26510 Fakulta podnikatelská VUT
  - 27510 Ekonomická fakulta VŠB [signif 3]
  - 28120 Fakulta managementu a ekonomiky UTBZ
  - 31110 Fakulta financí a účetnictví VŠE
  - 31120 Fakulta mezinárodních vztahů VŠE [signif 3]
  - 31130 Fakulta podnikohospodářská VŠE

  - 31160 Fakulta managementu v Jindřichově Hradci VŠE
  - 41110 Provozně ekonomická fakulta ČZU [signif 3]
  - 43110 Provozně ekonomická fakulta MENDELU
  - 7U000 Vysoká škola finanční a správní, a.s.
- ISCED-F 5 – Natural sciences, mathematics and statistics
  - 11310 PřF UK [signif 1 and 2]
  - 11320 MFF UK –[signif 6]
  - 12310 Přírodovědecká fakulta JUČB
  - 13440 Přf UJEP [signif 6 and 1]
  - 13520 FŽP UJEP
  - 14310 Přírodovědecká fakulta MU
  - 15310 Přírodovědecká fakulta UPO
  - 17310 Přírodovědecká fakulta OU [signif 6]
  - 18470 Přírodovědecká fakulta UHK [signif 1]
  - 22330 Fakulta potravinářské a biochemické technologie VŠCHT
  - 24510 Fakulta přírodovědně-humanitní a pedagogická TUL
  - 25310 Fakulta chemicko-technologická UP [7 signif too]
  - 26310 Fakulta chemická VUT
  - 31140 Fakulta informatiky a statistiky VŠE [signif 6]
  - 41330 Fakulta životního prostředí ČZU
- ISCED-F 6 – Information and Communication Technologies
  - 14330 Fakulta informatiky MU
  - 18450 Fakulta informatiky a managementu UHK
  - 21240 Fakulta informačních technologií ČVUT
  - 25410 Fakulta ekonomicko-správní UP [signif 3 and 4]
  - 26230 Fakulta informačních technologií VUT
- ISCED-F 7 – Engineering, manufacturing and construction
  - 13420 Fakulta strojního inženýrství UJEP
  - 21110 Fakulta stavební ČVUT
  - 21220 Fakulta strojní ČVUT
  - 21230 Fakulta elektrotechnická ČVUT [signif 6]
  - 21260 Fakulta dopravní ČVUT [signif 10 and 7]
  - 21340 Fakulta jaderná a fyzikálně inženýrská ČVUT [signif 5]
  - 21450 Fakulta architektury ČVUT
  - 22310 Fakulta chemické technologie VŠCHT [signif 5]

  - 22320 Fakulta technologie ochrany prostředí VŠCHT
  - 22340 Fakulta chemicko-inženýrská VŠCHT [signif 5]
  - 23210 Fakulta strojní ZUP
  - 23220 Fakulta elektrotechnická ZUP
  - 23520 Fakulta aplikovaných věd ZUP [signif 6 and 5]
  - 24210 Fakulta strojní TUL
  - 24220 Fakulta mechatroniky, informatiky a mezioborových studií ZUP [signif 06]
  - 24410 Fakulta textilní TUL
  - 25510 Dopravní fakulta Jana Pernera UP
  - 25530 Fakulta elektrotechniky a informatiky UP
  - 26110 Fakulta stavební VUT
  - 26210 Fakulta strojního inženýrství VUT
  - 26220 Fakulta elektrotechniky a komunikačních technologií VUT
  - 26410 Fakulta architektury VUT
  - 27120 Fakulta stavební VŠB
  - 27230 Fakulta strojní VŠB
  - 27240 Fakulta elektrotechniky a informatiky VŠB [signif 6 and 5]
  - 27350 Hornicko-geologická fakulta VŠB
  - 27360 Fakulta materiálově-technologická VŠB [signif 4]
  - 28110 Fakulta technologická UTBZ
  - 28140 Fakulta aplikované informatiky UTBZ [signif 6]
- ISCED-F 8 – Agriculture, forestry, fisheries and veterinary
  - 12220 Zemědělská fakulta JUČB [signif 7]
  - 12520 Fakulta rybářství a ochrany vod [signif 5]
  - 16170 Fakulta veterinárního lékařství VUB
  - 16270 Fakulta veterinární hygieny a ekologie VUB
  - 41210 Fak. agrobiologie, potrav. a přír. zdr. ČZU
  - 41310 Technická fakulta ČZU [signif 07]
  - 41320 Fakulta lesnická a dřevařská ČZU
  - 41340 Fakulta tropického zemědělství ČZU [signif 5]
  - 43210 Agronomická fakulta MENDELU [signif 5]
  - 43410 Lesnická a dřevařská fakulta MENDELU
  - 43510 Zahradnická fakulta (Lednice) MENDELU
- ISCED-F 9 – Health and welfare
  - 11100 1. lékařská UK

– 11200 2. lékařská UK
– 11300 3. lékařská UK
– 11400 Lékařská fakulta v Plzni
– 11500 Lékařská fakulta v Hradci Králové
– 11600 Farmaceutická fakulta Hradec Králové
– 12110 Zdravotně sociální fakulta
– 13450 FZS UJEP
– 14110 Lékařská fakulta MU
– 14160 Farmaceutická fakulta MU
– 15110 Lékařská fakulta UPO
– 15120 Fakulta zdravotnických věd
– 17110 Lékařská fakulta OU
– 17200 Fakulta sociálních studií OU
– 21460 Fakulta biomedicínského inženýrství ČVUT [signif 6]
– 25520 Fakulta zdravotnických studií UP