The Economics of
Knowing and Knowledge Creation

Inauguraldissertation zur Erlangung des Grades Doctor oeconomiae publicae (Dr. oec. publ.)
an der Volkswirtschaftlichen Fakultät der Ludwig-Maximilians-Universität München
vorgelegt von

Lena Anna Cäcilia Greska
2025

Referent:	Prof. Fabian Waldinger, Ph.D.
Koreferent:	Prof. Claudia Steinwender, Ph.D.
Promotionsabschlussberatung:	16. Juli 2025
Datum der mündlichen Prüfung:	30. Juni 2025
Namen der Berichterstatter:innen:	Fabian Waldinger, Claudia Steinwender, Joachim Winter

Acknowledgements

Newton's "standing on the shoulders of giants" is perhaps the most popular quote in the science of science. Far be it from me to not borrow the metaphor as well. I am very fortunate to have been able to lean on many giants during my PhD. My advisor Fabian Waldinger, who has helped me navigate many a complex econometric strategy and the myriad non-academic challenges of a PhD. Thank you for your encouragement and for teaching me so much. I also want to thank my second advisor, Claudia Steinwender for deeply insightful comments and perspicacious academic and life advice. Thank you to Joachim Winter for completing my dissertation committee, a great TA-ing experience, and many helpful comments.

Many others have also influenced my academic journey. Till Stowasser, thank you for "discovering me" during my Bachelor's at LMU, for teaching me to clean data like a hawk and for being my first ever coauthor. Yves Le Yaouanq, thank you for encouraging me to pursue a PhD and being my junior mentor. Carlo Schwarz, coauthor extraordinaire, thank you for being such a great collaborator and friend. Thank you also to Ran Abramitzky, Santiago Pérez and Joseph Price for an inspiring and fruitful transatlantic collaboration, including an amazing visit to Stanford. Pierre Azoulay enabled me to spend a transformative year at Massachusetts Institute of Technology, for which I am very grateful.

I owe many of the best memories from my studies to my colleagues and friends, Bernhard, Christina, Friederike, Jae, Kevin, Leonie, Maria, Paula, Peter, Robert, Robin, Sebastian, Silvia and Svenja. You have accompanied me through my bachelor's degree, the MQE, and endured my mid-morning chattiness as office mates. Together, we stuck out Covid, graded exams until the middle of the night and hashed out all the problems in our research. To my "interdisciplinary family", friends and PhD students at other LMU faculties – thank you for broadening my horizons over lunch and coffee. A special thanks goes out to Valentin Hoffmann, whose expertise in computational linguistics was fundamental in developing the specialization index.

I am also deeply indebted to the many institutions who made this research possible. The Egon Sohmen Graduate Center, the Joachim Herz Foundation, the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG) through CRC TRR 190 all provided financial support for the projects in this thesis.

The last five years have coincided with a difficult time in my life. I am eternally grateful to all the giants who carried me through it. Most of all, my relentlessly supportive family. You have built the foundations for this work by giving me curiosity, a thirst for learning, confidence in my abilities, and by inspiring my interest in generalists. Thank you for everything.

Contents

List of Figures

List of Tables

Preface

Knowledge fuels economic activity. At the micro-level, knowledge informs individual decisions. At the macro-level, knowledge creation accelerates technology-driven growth. This dissertation examines facets of knowing and knowledge in the economy in three independent chapters. All share a focus on the individuals who create knowledge, from specialized scientists solving problems in teams, scholars from lower socio-economic backgrounds working at U.S. universities, to partisan citizens making sense of the state of the economy. I explore questions ranging from how an individual's scope of knowledge affects innovation, to who creates knowledge in society, to when knowledge itself might be subjective.

Chapter 1 examines the trade-off between specialization and coordination costs in innovative teamwork. The last century has seen a steady rise in scientific specialization (Jones, 2009), coinciding with an increase in scientific teamwork (Wuchty et al., 2007). Specialization enhances scientific productivity by allowing scientists to hone their skills and to profit from the division of labor. However, teams must coordinate. Highly specialized scientists may struggle to do so, for instance because they cannot adapt to the work of teammates with different skill sets. I formalize this trade-off in a model of scientific collaboration between generalists and specialists and test the predictions of the model in the context of online machine learning competitions. Using a novel measure of specialization based on the semantic diversity of code, I find that generalist teams produce higher quality innovation than specialist teams when coordination costs are high. This has implications for the design of scientific processes for both firms staffing research and development teams and for universities and science funding agencies looking to increase interdisciplinary collaboration. While previous research suggests that teams from more diverse backgrounds produce higher-impact innovation (Uzzi et al., 2013), I highlight that more demanding coordination in these teams could lead to the failure of such endeavors.

Chapter 2, co-authored with Ran Abramitzky, Santiago Pérez, Joseph Price, Carlo Schwarz, and Fabian Waldinger, focuses on a different aspect of diversity in knowledge creation: the representation of individuals from lower socio-economic backgrounds in academia. We assemble an unprecedentedly large dataset of U.S. academics and their backgrounds, spanning seven decades from 1900 to 1969. Throughout this time, individuals from poorer backgrounds have been severely and consistently underrepresented, despite a coinciding expansion in U.S. (higher) education. Socio-economic background is strongly intertwined with where knowledge is created, what knowledge is created, and whether knowledge gets recognized. Humanities and elite universities are the purview of academics from richer backgrounds. Academic disciplines often mirror paternal occupations; for

example, children of physicians become professors of medicine. Despite similar publication records introducing more novel concepts, academics from poorer backgrounds receive fewer citations and are less likely to be nominated for a Nobel Prize than their richer counterparts.

Chapter 3, co-authored with Till Stowasser, investigates *knowing*. While the previous chapters have taken knowledge creation as the process of discovering truths about empirical facts, we now examine knowing as a subjective process. We show that supporters of different parties perceive the economy differently depending on who is in power. We argue that this is driven by motivated information processing. Partisans prefer to see "their" party in a good light, distorting the processing of information about economic conditions to fit their beliefs. We develop a novel methodology, *synthetic beliefs*, to detect such phenomena from aggregated survey data.

Chapter 1

Innovative Collaboration: Generalists, Specialists, and Teams

---

This chapter is based on single-authored work. I gratefully acknowledge financial support from Joachim Herz Foundation and the German Research Foundation (DFG) through CRC-TRR 190. Krisha Agrawal provided excellent research assistance.

1.1 Introduction

Economists have long recognized specialization as fundamental to prosperity. In the first chapter of *The Wealth of Nations*, "Of the division of labour", Adam Smith argued that "The greatest improvements in the productive powers of labour [...] seem to have been the effects of the division of labour." (Smith, 1776). Similar arguments for specialization have been made in science and innovation. Here, the need to specialize is particularly pressing, since innovation builds on prior knowledge, and a single person cannot know everything (Jones, 2009). The well-documented rise in scientific specialization over the last century (Jones, 2009) has coincided with an increase in scientific teamwork (Wuchty et al., 2007).

Yet, teamwork and specialization are hard to reconcile in practice: Efficient teamwork requires coordination. Coordination is more demanding when individual team members are very specialized. For example, it becomes harder to understand and adapt to others' work, integrate individual contributions into a cohesive whole, and provide feedback. In this paper, I investigate the trade-off between specialization and coordination in scientific teamwork. I show theoretically that a team of generalists – scientists with low levels of specialization – produces the highest quality innovation when coordination costs are high. I then test the predictions of the model in the context of online machine learning competitions. Employing various econometric strategies to isolate the team type effect from the effect of individual team members, I find that generalist teams develop higher quality innovation than specialist teams. I present a range of evidence which points to coordination costs as the mechanism that drives generalist-specialist quality differences rather than other factors.

In my theoretical framework, scientists with heterogeneous levels of specialization work alone or collaborate to solve problems with varying complexity. Similar to Becker and Murphy (1992), on which my model builds, each problem requires a set of tasks to be executed in order to solve it. The quality of the problem's solution depends on the quality of execution of these sub-tasks. In a blend of Deming (2017) and Garicano (2000), scientists' specialization is modeled as different ex-ante task-specific productivities. A generalist, with zero specialization, can execute all tasks but each with lower productivity. The more specialized a scientist, the fewer tasks they can solve, but the higher their productivity for these tasks. Here, I depart from Becker and Murphy (1992), where all team members are equally specialized ex-post as a result of dividing tasks in a team. When collaborating, scientists can split the tasks necessary to solve a problem, and produce quality more efficiently. However, they incur coordination costs, which reduce the quality of their solutions. Importantly, coordination costs increase (1) in each team member's degree of specialization, and (2) in the complexity of the problem they are trying to solve. In this aspect, I deviate from other models of scientific teamwork and specialization, which often assume costless aggregation of knowledge in teams (e.g., Jones, 2009).

I derive two predictions from this framework: First, if coordination costs are sufficiently high, a team of generalists solves a problem with the highest quality. Thus, a decrease in coordination costs will lower or eliminate quality differences between generalists and specialists. Second, complexity increases quality differences between generalist and specialist teams if coordination costs are sufficiently high. These predictions allow me to develop empirical tests to pin down coordination costs as the mechanism behind observed quality differences in teams' innovative output.

Testing these predictions requires an empirical setting that allows me to observe how well different types of teams solve problems *holding the problem constant*, and how well different problems get solved, *holding team types constant*. To compare solutions, I need to be able to construct an objective measure of solution quality. To accurately categorize teams, I need to quantify each team member's specialization. Crucially, the setting needs to allow me to isolate the effect of individual team members' ability from the effect of team type on quality. Online machine learning competitions provide an optimal environment for studying specialized scientific teamwork as they satisfy all these criteria. With a long tradition in computer science, specifically in machine learning and artificial intelligence (Koch and Peterson, 2024), the objective of these competitions is to develop an algorithm that solves a given, well-specified problem. The resulting algorithm's quality is evaluated according to a pre-specified, objective performance metric. I use data from Kaggle.com (Kaggle), a large online machine learning platform which provides infrastructure for (research) institutions and corporations to host machine learning competitions.[1] Kaggle users are computer science professionals like artificial intelligence researchers, machine learning engineers, and data scientists. Kaggle's features are particularly suitable for my study: First, problems – competitions – are well-defined, and have a large number of teams and solo competitors trying to solve them. Second, the quality of solutions is measured in an objective way for all teams attempting to solve the problem: The host of the competition specifies a quality metric, and all submitted solutions are ranked according to this criterion. Third, Kaggle users also publish code on the platform, from which I construct my measure of specialization. Fourth, users frequently work alone. This enables me to observe individual performance independently of team performance, forming the foundation of my three-pronged empirical strategy: Explicitly controlling for observable heterogeneity between specialists and generalists, employing user-fixed effects to account for time-invariant unobservables, and conducting an event study based on team formation timing.

I develop a new measure of specialization for the computer sciences based on the breadth of a computer scientist's code portfolio. Intuitively, to solve a computer science problem, a computer scientist needs to write code. Broader coding skills reflect an ability to tackle a broader range of problems, and hence, less specialization. To quantify the breadth of coding skills, I use CodeT5+

[1] As of November 2024, Kaggle has more than 21 million global users and hosted 407 competitions awarding a total of USD 22.3 million in prizes.

(Wang et al., 2023), an open-source large language model (LLM), trained on a substantial corpus of code and comments from GitHub. With CodeT5+, I transform code into "embeddings" – vector representations capturing the semantic meaning of code. Embeddings allow me to quantitatively compare different code snippets by calculating cosine similarity, a standard metric from natural language processing used to assess the conceptual similarity of text. My specialization measure then is the average cosine similarity between all pairs of code instances written by a given scientist. A high average similarity indicates that the scientist's code tends to cluster within a narrow area, suggesting specialization. Conversely, a low average similarity implies the scientist works across a broader range of coding domains. Scientists with above-median similarity scores (i.e., lower breadth) are classified as specialists, while those below the median are classified as generalists.

Using my measure of specialization and quality measures from Kaggle, I show that generalist teams achieve higher quality solutions than specialist or mixed teams. In my preferred specification, a solution by a team comprised solely of generalists is ranked 4 percentiles higher than a solution by a specialist team, corresponding to placing on average 50 absolute ranks higher. In addition, generalist teams are 8 percentage points more likely to win a medal, a prestigious labor market signal, and 5 percentage points more likely to reach the coveted top three positions in a competition associated with large monetary prices. Mixed teams place between generalist and specialist teams.

I employ a range of empirical strategies that support the interpretation of these quality differences as fundamental differences between specialist and generalist teams, over and above differences between individual generalists and specialists. First, I control for a large set of individual characteristics to capture any observable heterogeneities other than specialization that might influence solution quality, like experience or ability. Second, to account for time-invariant unobservable scientist characteristics that might differ between generalists and specialists, I estimate a regression with user fixed effects. Third, to address concerns that performance in other competitions is not a good counterfactual for performance at a given task, I make use of a unique feature of Kaggle: Competitors can submit preliminary solutions during a competition to get a signal of the quality of their solution. Since some teams only form during the competition, I implement an event study design similar to Lemus and Marshall (2024) to assess how the quality of preliminary solutions changes before and after team formation. Generalist teams outperform specialist teams in all empirical approaches.

After establishing solution quality differences between generalist and specialist teams, I turn to testing the mechanism that creates these differences in the model, coordination costs. Since I cannot measure coordination costs directly, I use two shifters of coordination costs to assess their impact on generalist and specialist teams. First, I split the sample into high and low complexity competitions. I can only detect quality differences in high complexity competitions, indicating that

complexity indeed mediates coordination costs as suggested by my theoretical model. Second, I use the introduction of ChatGPT as an exogenous shock to coordination costs. ChatGPT can smooth some of the frictions in specialized collaboration, for instance by explaining unfamiliar concepts and the function of unfamiliar code in particular, translating one coding language into another, or writing code to integrate disjoined components of an algorithm. Again, I split the sample into competitions before and after the introduction of ChatGPT, and only find quality differences in the period prior to the introduction of ChatGPT.

I also examine a range of alternative mechanisms suggested by the previous literature on teamwork (e.g., Deming, 2017; Weidmann and Deming, 2021; Ahmadpoor and Jones, 2019; Adhvaryu et al., 2023; Freund, 2024; Minni, 2024; Weidmann et al., 2024): higher effort by generalist teams, better management in generalist teams, higher social skills in generalist teams, and better collaborator matching in generalist teams. I do not find any evidence in support of these alternative mechanisms. Taken together, my findings indicate that coordination costs indeed reduce the productivity of specialist collaboration in innovation.

This paper contributes to the literature on specialization and teamwork in science and innovation. A large body of research has documented an increasing dominance of teams in innovation (Wuchty et al., 2007; Jones, 2009; Ahmadpoor and Jones, 2019; Pearce, 2023) as well as increasing specialization (Jones, 2009; Agrawal et al., 2016). The implications of these two trends for scientific productivity have often been examined through the lens of "ideas being harder to find" (Bloom et al., 2020), specifically, that implementing ideas increasingly requires differentiated expertise. Allocca (2024) suggests that teams with more diverse educational backgrounds are more likely to successfully complete research projects. Pearce (2023) finds that patents by research teams with more diverse expertise receive more citations. While these studies consider the diversity of expertise embodied by the team, I focus on the diversity of expertise embodied by one individual, their *specialization*, and how this impacts their productivity in scientific teamwork. This dimension has been less examined by the literature. Notable exceptions are Teodoridis (2018), who documents changes in the rate of generalist-specialist collaboration after a technology shock that substituted for specialist skills, and Teodoridis et al. (2019), who provide evidence for individual generalists outperforming specialists when the overall rate of innovation is slower. To the best of my knowledge, there is no paper that empirically examines the effect of generalist-specialist team composition on innovative productivity.

My paper also contributes to the largely theoretical literature on the trade-off between specialization and coordination in teamwork in organizational economics. Starting with Becker and Murphy (1992), multiple studies show how decreases in coordination costs, sometimes modeled as communication costs (e.g., Bolton and Dewatripont, 1994), allow for higher levels of profitable specialization

in teams and hierarchies (e.g., Garicano, 2000; Garicano and Rossi-Hansberg, 2006). Conversely, Dessein and Santos (2006) show how decreases in communication costs can lead to *lower* levels of specialization within an organization by allowing the organization to become more adaptive to new information. However, none of these incorporate heterogeneous levels of specialization. Recent work by Freund (2024) models specialization as ex-ante differences in the dispersion of task-specific productivity, similar to my framework, but there are no coordination frictions within the team. Perhaps closest in spirit to my theoretical framework is Deming (2017), where workers with different task-specific skills coordinate on "trading" tasks to produce a final good. The efficiency of this trade is determined by social skills, which are orthogonal to task-specific skills, i.e., specialization. I endogenize coordination costs to a team member's specialization to more explicitly capture the trade-off between specialization and coordination.

I also contribute to the literature on what makes teams succeed. Recent contributions have highlighted social skills (Deming, 2017; Weidmann and Deming, 2021; Adhvaryu et al., 2023), collaboration experience and team-specific human capital (Jaravel et al., 2018; Chen, 2021), and leadership (Minni, 2024). I add generalists to this list. Finally, a small body of research has explored teamwork and innovation on Kaggle and similar platforms, with a specific focus on tournament effects (Boudreau et al., 2011, 2016; Lemus and Marshall, 2021, 2024). For example, Lemus and Marshall (2024) find that teamwork improves performance compared to solo competitors. I take this finding further: How can we shape teams to generate maximum innovation?

1.2 A Model of Collaboration between Generalists and Specialists

In this section, I develop a model of scientific collaboration between specialized scientists. Scientists with heterogeneous levels of specialization solve problems alone or in teams. Specialization affects scientists' problem solving in three ways: Which problems a scientist can solve, how well scientists can solve a problem, and how well scientists collaborate. I derive two testable predictions from this framework. To keep notation and exposition as simple as possible, the focus of my analysis is on teams of two scientists.

1.2.1 Set-Up

Problems Scientists solve problems p. To solve a problem, scientists have to execute a continuum of tasks Θ_p. Like Becker and Murphy (1992), I assume tasks are perfect complements, where the problem's overall solution quality is determined by the quality of the worst-executed task. While stringent, this assumption keeps the model tractable and is not too far from reality. Consider, for

instance, conducting empirical research as consisting of two tasks, data cleaning and regression analysis. If the data are not cleaned properly, no amount of well-specified regressions can lead to sound conclusions. Conversely, even perfect data will not save a misspecified regression. Problems differ along two dimensions, (1) the amount $\Theta_p \in (0, 1]$ of different tasks that need to be executed to solve the problem and (2) complexity γ_p. I formalize complexity as a factor that makes teamwork more difficult.

Scientists Scientists i have different levels of specialization σ_i. Specialization has three effects on a scientist's problem solving ability. First, it determines how many different tasks a scientist can execute, where $b_i \in (0, 1]$ is the range of tasks the scientist can execute. The range of tasks a scientist can solve is inversely proportional to their specialization, i.e., the higher a scientist's level of specialization, the fewer tasks the scientist can execute ($\sigma_i = \frac{1}{b_i}$). For expositional simplicity, I consider a scientist who can execute all tasks ($\sigma_i = b_i = 1$) a generalist. Specialization also affects how well a scientist can execute a task. Conditional on being able to execute a given task, a scientist with a higher level of specialization executes the task with higher quality. These first two effects are similar to the most prominent ways of modeling specialization in the literature – heterogeneous task-specific productivity (Deming, 2017; Freund, 2024) and whether or not a task can be executed (Garicano, 2000; Jones, 2009). The third effect is novel and where the main contribution of my model lies. Specialization also determines how easy it is for a scientist to collaborate. Teamwork is more difficult if scientists are very specialized. In the empirical research example, imagine two scientists are collaborating on a study, but one scientist only knows how to clean data, and the other only how to run regressions. Coordinating on how to conduct the study would be extremely difficult in this case.

Collaboration Scientists can choose whether to try and solve a problem on their own or whether to collaborate. For simplicity, I abstract from collaborator search. Instead, I consider a world in which, for each problem, a scientist is matched at random with another scientist, and their only choice is whether to collaborate or not. When collaborating, the two scientists have to split the Θ_p tasks among themselves. Let w_i denote the share of tasks executed by scientist i. When collaborating, the team of scientists incurs coordination costs C. Coordination costs capture all the frictions that arise in team production but not when working alone, for example, deciding and negotiating how to split tasks, which tasks to bundle, or determining how to aggregate the output from sub-tasks into the final product.

Problem Solving and Payoffs A problem p is solved by scientist i with quality Q_p:

$$Q_p = \min_{0 \leq \theta \leq \Theta} q(\theta) \qquad \text{where} \quad q(\theta) = \begin{cases} 0 & \text{if } \theta \notin b_i \\ q(w_i, \sigma_i, e_i) & \text{if } \theta \in b_i \end{cases} \tag{1.1}$$

e_i is the effort spent by scientist i on the θ^{th} task. The quality of execution of each task thus depends on the amount of tasks that one scientist needs to execute w_i, a scientist's specialization σ_i, and the effort spent on each task e_i. Intuitively, $\partial q(w_i, \sigma_i, e_i)/\partial e_i \partial \sigma_i > 0$, meaning that conditional on being able to execute a task, a more specialized scientist has a higher marginal productivity of effort. The marginal productivity of effort however decreases in the amount of tasks a scientist has to execute, i.e., $\partial q(w_i, \sigma_i, e_i)/\partial e_i \partial w_i < 0$. Similar to assumptions made by Deming (2017) and Becker and Murphy (1992), this allows scientists to become more productive as they focus on fewer tasks, and is the main mechanism through which teamwork can be beneficial. An intuitive interpretation is that when scientists have to execute many task, they have to split their attention and get distracted more easily, or cannot learn as much about the individual task.

Scientists maximize payoffs, equal to a problem's solution quality minus the total cost of effort $w_i c(e_i)$, where $c(e_i)$ is the effort spent on one task θ. Since tasks are symmetric and perfect complements, the total cost of effort is the cost of effort for each task multiplied by the share of tasks executed by one scientist, w_i. That is, when working alone, scientist's payoffs Π_p^a from solving problem p are given by:

$$\Pi_p^a = \begin{cases} 0 & \text{if } \Theta_p > b_i \\ \min_{0 \leq \theta \leq \Theta_p} q(\Theta_p, \sigma_i, e_i) - \Theta_p c(e_i) & \text{if } \Theta_p \leq b_i \end{cases} \tag{1.2}$$

where $\Theta_p c(e_i)$ is the cost of effort of executing all Θ_p tasks required to solve a problem. Scientists receive utility from solving a problem well, such that the quality of a problem's solution maps directly into scientists' payoffs. When collaborating, scientists' payoffs are equal to *half* the problem's solution quality. However, scientists can achieve quality more efficiently because they can split tasks to increase their marginal productivity of effort. They can also potentially collaborate with a more productive, i.e., more specialized, scientist then themselves. Splitting tasks also means that scientists have lower total effort costs for the same task-specific effort e_i, since they execute fewer tasks. Yet, teams also incur coordination costs. A scientist i's payoffs Π_p^t from solving problem p in a team with scientist j with specialization σ_j are given by:

$$\Pi_p^t = \frac{1}{2}\{\min_{0 \leq \theta \leq \Theta_p} q(w_i, \Theta_p - w_i, \sigma_i, \sigma_j, e_i, e_j) - C(\gamma_p, \sigma_i, \sigma_j)\} - w_i c(e_i) \tag{1.3}$$

where e_j is the effort executed by scientist j. That is, payoffs crucially depend on the other team member. Coordination costs $C(\gamma_p, \sigma_i, \sigma_j)$ reduce a problem's solution quality and hence payoffs for both team members, and are an increasing function of the problem's complexity γ_p, and both team members' degree of specialization, i.e., $\partial C/\partial \gamma_p > 0$ and $\partial C/\partial \sigma_i > 0$. Importantly, complexity has a stronger impact on more specialized scientists, i.e., $\partial C/\partial \gamma_p \partial \sigma_i > 0$. Why is that? Going back to the previous example, imagine the two scientists now have to implement an econometric strategy that requires careful data preparation, for example, in matching control observations on a complex range of characteristics. If the team member who cleans the data has no understanding of the identification challenges, they might focus on the wrong characteristics and produce a matched sample that exacerbates bias. The same complexity is less harmful if the team member who cleans the data also understands identification.

1.2.1.1 Choices

Scientists make three main choices. First, they choose whether to collaborate or not. Second, only if collaborating, they choose how to allocate tasks. Finally, they choose the optimal level of effort.

Collaboration Choice When choosing to collaborate, scientists trade off the gains from the division of tasks with collaboration costs. Two scientists form a team when the payoffs to teamwork are larger than the payoffs to working alone for both scientists in the team.

$$\frac{1}{2}\underbrace{\{Q_p^t - C(\gamma_p, \sigma_i, \sigma_j)\}}_{\text{team solution quality}} - \underbrace{w_i c(e_i^{t*})}_{\text{total cost of effort}} \geq \underbrace{Q_p^a}_{\text{individual solution quality}} - \underbrace{\Theta_p c(e_i^{a*})}_{\text{total cost of effort}} \tag{1.4}$$

If scientists choose not to collaborate, they either solve the problem on their own if $\Theta_p \leq b_i$ or do not participate if $\Theta_p > b_i$.

Task Division and Effort Choice Scientists choose the division of task to maximize solution quality minus total effort costs. Optimal task division and efforts are given by:

$$w_i^* = \begin{cases} \Theta_p & \text{if } n = 1, \\ \arg\max\limits_{w_i} \quad \frac{1}{2}\min q(w_i, \Theta_p - w_i, \sigma_i, \sigma_j, e_i^*, e_j^*) - w_i c(e_i^*) & \text{if } n = 2. \end{cases} \tag{1.5}$$

and

$$e_i^* = \arg\max_{e_i} \quad \frac{1}{n}\min q(e_i, \sigma_i, w_i) - w_i c(e_i) \tag{1.6}$$

where n is 1 when working alone and 2 when collaborating. When working alone, w_i is Θ_p by definition. In a team, the optimal division of tasks will be pinned down by the fact that for both scientists, marginal productivity has to equal the marginal change to total effort costs for both scientists, and that each scientist has to produce the exact same quality in each sub-task.

1.2.2 Predictions

In the model, scientists trade off higher productivity in specialized team work and coordination costs. I derive two predictions from this framework to test whether coordination costs really do limit the returns from specialization. For more detailed derivations, see Appendix 1.A.

Prediction 1. *If coordination costs are sufficiently high, a team of generalists produces the highest quality solution. Conversely, a decrease in coordination costs will reduce quality differences between generalist and specialist teams.*

Proof: This follows directly from applying the envelope theorem to the function of team solution quality.

Prediction 2. *An increase in complexity will increase the quality differences between generalist and specialist teams.*

Proof: Since $\partial C/\partial\gamma_p\partial\sigma > 0$ quality differences between teams will be larger when Prediction 1 holds.

1.3 Machine Learning Competitions and Kaggle

1.3.1 Why machine learning competitions?

In machine learning, competitions are widely used both to accelerate innovation and to evaluate the quality of innovation. Competitions were first introduced by the Defense Advanced Research Projects Agency (DARPA) in the 1980s to allocate grants for risky artificial intelligence projects (Koch and Peterson, 2024) and are still widely used (Pavão, 2023).[2] A prominent example is the 2006 Netflix Prize, in which the streaming platform Netflix awarded USD 1 million to the team that would achieve 10% higher prediction accuracy than Netflix's proprietary movie recommendation algorithm (Bennet and Lanning, 2007).[3] Competitions are also an integral part of academic computer science. For example, prestigious computer science conferences like Knowledge Discovery and Data Mining

[2]Theory and empirical evidence also indicate that competitions are an efficient method to generate innovation (Lazear and Rosen, 1981; Boudreau et al., 2011, 2016; Kireyev, 2020).

[3]Note that a competition is distinct from a hackathon. While hackathons are more open-ended ideation challenges within a thematic complex, competitions are directed innovation focusing on a single computation problem.

(KDD) and Neural Information Processing Systems (NeurIPS) include a dedicated competition track (Albrecht et al., 2024; Association for Computing Machinery, 2025). The basic structure of any machine learning competition has remained the same since their introduction: The host of the competition describes the problem at hand, provides the dataset as input for the desired algorithm, and sets timeline, evaluation metric, and prizes. Crucially, only a subset of the dataset (*training data*) is made available to competitors to develop their algorithm. Once the competition ends, all competitors submit an algorithm. The performance of the submitted algorithm is evaluated against a dataset that was not previously known to competitors (*test data*). Whosever algorithm achieves the best performance in the test data wins the competition.

Machine learning competitions are an attractive setting for the study of innovative teamwork for several reasons. First, problems are well-defined. Second, performance is measured in a unified, objective way: ranks by performance of the algorithm.[4] This enables me to compare the performance of different team types at the exact same problem. Third, online competition platforms like Kaggle provide rich data that facilitate econometric analysis. In any team study it is necessary to isolate the effect of the team type from the characteristics of the team members (Weidmann and Deming, 2021; Bonhomme, 2022). This requires observing performance of team members when they are working alone. For many science and innovation settings this poses a challenge, since both patents and papers are increasingly produced in teams (Wuchty et al., 2007). On Kaggle, only 1% of all users never work alone, such that I can observe individual performance for almost all team members. Additionally, several features of the platform allow me to construct measures for concepts like effort and social skills to explore mechanisms. Most importantly, I can measure specialization.

1.3.2 Kaggle

Kaggle is a platform for data science and machine learning. Founded in 2012, its core business is providing infrastructure for companies and research institutions to host machine learning competitions, but it also serves as a learning, networking, as well as data, models and code sharing platform for computer science professionals. As of November 2024, Kaggle has more than 21 million global users and hosted 407 competitions awarding a total of USD 22.3 million in prizes. Kaggle makes all data from their users' public activities available on their website as a downloadable dataset, Meta Kaggle (Risdal and Bozsolik, 2022). The dataset contains granular information on users, competitions, teams, discussions, and code meta data. The main focus of this paper are competitions. However, I utilize users' activities in the other areas to explore mechanisms and construct main variables.

[4]This compares to much of the academic literature in computer science: Papers benchmark their proposed algorithms to the state of the art.

Competitions My analysis focuses on the two most prominent types of competitions on Kaggle, *Featured* and *Research* competitions. Of all users who ever participated in any competition, 46.8% participated in a featured competition and 15.2% in a research competition.[5] Though mostly similar, featured and research competitions differ slightly in terms of the type of problems they cover – featured competitions tend to focus more on commercially applicable problems, whereas research competitions focus on scholarly and scientific problems.[6] The term competitions will refer to *Featured* and *Research* competitions from now on. Kaggle competitions proceed like typical machine learning competitions. A competition host – a company or a research institute – sets up a competition. The host describes the goal of the competition (e.g., detecting cancer from mammograms, recognizing bird voices from audio), provides labeled data, determines the evaluation metric[7], the prizes, and the competition timeline. Competitions run for three months on average, and there is an entry and team formation deadline. Appendix Table 1.B.1 lists example competitions.

Once a competition begins, users are invited to work on the problem and submit their solutions. A feature particular to Kaggle is that users can obtain preliminary feedback on the performance of their algorithm in real time. Kaggle splits the data provided by the competition host into three datasets: The training set is available to users to train their algorithms, the *public* test set is used to evaluate submissions during the competition, and the *private* test set is used to evaluate submissions once the competition ends. This split is meant to ensure that competitors get feedback on their algorithms during the competition, but that solutions are actually useful by reducing the incentive to overfit to the available data. Competitors' ranks in terms of the score of their solution are displayed publicly on the *leaderboard* in real time during the competition.[8] At the end of the competition, users can select two submissions to be judged against the private test set. Submission scores at this state determine the winners of the competition. I observe public and private ranks of all submissions made to any competition.

Teams Competition participants can choose to work alone or to collaborate with up to four other competitors. Competitors first enter a competition on their own and can then form or join a team at any point during the competition up to a predetermined team formation deadline. Teams can only

[5]All numbers refer to the version of Meta Kaggle used in this paper, downloaded on February 7, 2024 (Risdal and Bozsolik, 2022).

[6]I exclude the other types of competitions, *Community*, *Playground*, and *Getting Started*, from the analysis since the problems posed in these competitions are more akin to classroom problems, and performance is not incentivized. I also exclude *Recruitment* competitions, used by companies like AirBnB or Facebook as assessment centers for data scientists, in which teamwork is prohibited.

[7]Usually, these are prediction accuracy metrics commonly used in machine learning, like Root Mean Square Error or Log Loss. Some competitions additionally set infrastructure constraints, like run-time, maximum central processing unit (CPU) and graphics processing unit (GPU) usage, or internet access.

[8]For a discussion of the impact of intermediate feedback in tournaments and specifically Kaggle, see Lemus and Marshall (2021). They find that intermediate feedback generates better performing algorithms.

be dissolved if the team has not made any submission. If a team wins a competition prize, the prize is divided equally among all team members. Team formation is voluntary and self-organized. For example, Kaggle users can use public message boards attached to each competition to announce they are looking for teammates[9], or use Kaggle's messaging features to contact potential collaborators. Since Kaggle displays a range of information on each user's profile, such as in which competitions a user participated in the past, and how they performed, team members' abilities are public knowledge. Anecdotal evidence from interviews I conducted with active Kaggle users suggests that these observable performance signals determine users' team formation decisions. I observe each team member's past performance, as well as team sizes and when the team was formed.

Incentives Do teams actually try to develop the best algorithm? A common worry in the contest design literature is that a paucity of monetary prizes decreases incentives of all contestants, since each contestant has a low chance of winning (e.g., Taylor, 1995). Kaggle counteracts this with providing additional reputation incentives. Each user receives a platform-wide ranking, which is a function of competition performance, and belongs to a status group. Akin to chess, the most prestigious groups are masters and grandmasters. These can be achieved by performing extremely well, but not necessarily winning a monetary prize, in competitions. Both platform ranks and status groups are displayed on a user's Kaggle profile page and are valuable signals in the labor market for computer scientists. Users frequently include this information in their résumés (Reich, 2023), and several companies explicitly hire based on Kaggle rankings and status groups. For example, GPU manufacturer NVIDIA employs a team of Kaggle grandmasters solely to participate in Kaggle competitions (NVIDIA, 2025). Finally, competitors are also intrinsically motivated to contribute to open science. Although winners are not required to do so, many share their approaches and solutions after the competition. Kaggle competitions have resulted in research papers (e.g., Bulten et al., 2022; He et al., 2024), with competition winners as co-authors. Kaggle users also stress the innovation aspect of their work on Kaggle: "Normally, Kaggle is always one step ahead of science. [...] To perform really well in the competitions, one has to go an extra mile. Anyone can read what already exists. You have to come up with something new to beat 1000 other people."[10]

[9] In August 2022 Kaggle additionally introduced a waving-hand icon that users and teams who are looking for teammates can display next to their name on the leaderboard.

[10] Christof Henkel (alias Dieter), current highest-ranking Kaggle user and a member of the Kaggle Grandmasters of NVIDIA (Kaggle.com, 2024c), in zoom interview with the author, April 2024.

1.3.3 Measuring Specialization in Kaggle

1.3.3.1 Conceptual Considerations

A scientist's degree of specialization is equal to their breadth of knowledge. I thus need to find a representation of the knowledge held by one scientist. In computer science, this is what sort of code someone is able to write. For example, a computer scientist who writes code for both image data and text data has broader knowledge than a scientist who only ever writes code for text data. The first scientist is less specialized than the second. The main challenge is to quantify this distinction. Recent advances in computational linguistics make this possible through embedding models for code. These models, which also power large language models like ChatGPT, transform code into high-dimensional vectors that capture the semantic meaning of code. By learning the context in which instances of codes occur, embedding models are able to represent the relationship between two sets of code, as well as between code and plain text in an ordinal way. Two related concepts will be represented by two closely located vectors. Conversely, the distance between two vectors measures how different two concepts are. The totality of embeddings thus effectively maps the "knowledge space" of coding expertise. This feature makes embedding models ideally suited for the study of specialization.

To illustrate why, it is helpful to contrast this approach with existing measures in the literature. For example, Jones (2009) measures specialization of inventors as the probability of changing patent categories between to successive inventions. Teodoridis (2018) and Teodoridis et al. (2019) construct a diversification index based on the Herfindhal index of a scientist's paper categories. Discrete categories however fail to capture both the relationship of knowledge representations between and within categories. Imagine there are three types of code: `areg`, `xtreg`, and `hpfilter`. A category-based index would assign the same level of specialization to a user that has written one `xtreg` and `areg` code each, and one that has written one `xtreg` and one `hpfilter` code, although `xtreg` and `areg` are more closely related than `xtreg` and `hpfilter`. An embedding model resolves this issue – the distance between `areg` and `xtreg` will be smaller than the distance between `hpfilter` and `areg` or `xtreg`, respectively. An additional advantage of an embedding-based index is that it does not rely on labeled data, which is often costly or impossible to obtain. It is particularly infeasible in this context where code files are rarely labeled, although Kaggle enables authors to self-assign category tags to code files. Tag coverage is sparse, endogenous, and tags are often not meaningful (e.g., "advanced").

To quantify the "distance" between two instances of code, I use cosine similarity, a standard measure of textual similarity in computational linguistics. For each scientist, I then define specialization as the average cosine similarity between their individual code files and their core coding

discipline, represented by the centroid of a scientist's code files:

$$Specialization_i = \frac{1}{K_i} \sum_{1}^{K_i} cos(v_{ki}, c_i)$$

where K_i is the total number of code files k_i written by a user i, v_{ki} is the embedding of code file k_i, and c_i is the centroid (mean) embedding of user i, constructed as the average of all of user i's code file embeddings: $c_i = \frac{1}{K_i} \sum_{1}^{K_i} v_{ki}$. In less technical terms, this measure is analogous to the variance of knowledge around an individual's core discipline, represented by the centroid embedding.[11]

1.3.3.2 Construction of the Specialization Index in Kaggle

In addition to competitions, Kaggle provides infrastructure for running, publishing, and sharing code. Similar to competitions, users are incentivized to publish code by receiving medals, titles, and a platform-wide ranking.[12] Kaggle makes the entirety of shared code available as a corpus (Plotts and Risdal, 2023). In the corpus, code has several formats and languages, either unstructured files containing the entire code with comments, or a notebook-style format in which code, comments, and other exhibits are logically separated into cell blocks[13], which makes an additional processing step necessary. Before I transform each instance of code into an embedding, I remove all code from the data that has multiple authors, that is an edited version of code written by another user, has less than 10 lines, or is a duplicate. When the original author of a duplicate code file is not credited by the author of the copy, I manually search for the original author, assign the code file to only them and drop all duplicates. For each code file, I only keep the most current version. I also drop all code written by users who have published only one code file on the platform to ensure I have enough observations per user to reliably classify specialization.

I extract code embeddings from CodeT5+ (Wang et al., 2023). CodeT5+ is an open source large language model trained for several tasks, such as suggesting code to solve a specific problem, explain code, or complete code. CodeT5+ is particularly well-suited for my study because it is trained on both code and natural language comments, as well as entire programs rather than isolated code snippets. These features make it more effective for capturing broader programming contexts. Although I can technically separate code and comment data in some notebooks, it is non-trivial to remove all natural language from code files.[14] Including comments in the measure of specialization

[11] Taking the centroid as a base is necessary for an intuitive reason: Imagine two scientists starting at the same point in the knowledge space. Each scientist takes ten steps of equal length in this knowledge space. However, one scientist walks in a circle around the starting point, while the other walks ten steps away from the starting point. We would probably all consider the first scientist more specialized than the second.

[12] Although being a competition grandmaster is considered the more prestigious title, code grandmasters also provide this information on their LinkedIn, and are hired by leading AI companies, see, e.g., h2o.ai (2025).

[13] See Appendix Figure 1.B.1 for a visualization

[14] For example, code might be commented out, or a comment might follow a line of code. Additionally, comments are marked differently in different coding languages.

is also beneficial because it adds information on the task at hand. While CodeT5+ is able to process relatively large sequences of code and comments, many Kaggle code files exceed CodeT5+'s maximum sequence length[15]. To handle longer files, I split the code file into logical units, equivalent to paragraphs (see Figure 1.B.1 for an illustration), which I transform into embeddings and then aggregate.[16]

Once I have obtained embeddings for each code file, I construct the centroid (mean) embedding for each user, calculate the pairwise cosine similarity of each individual code embedding to the centroid, and take the average. This yields a continuous measure of specialization for 65.211 users, see Figure 1.1, Panel (a). A value closer to one indicates that all code written by a user is very similar, i.e., a higher level of specialization. However, mechanically, the distribution of the specialization index is shifted to the right when it is constructed from fewer code files, see Figure 1.1, Panel (b). To not confound the number of code files written with specialization, I adjust the index by residualizing it to the number of kernels, as shown in Panels (c) and (d). I then define *Generalists* as users with below median levels of specialization, and *Specialists* as users with above median levels of specialization.

1.3.4 Variable Definitions

1.3.4.1 Key Concepts

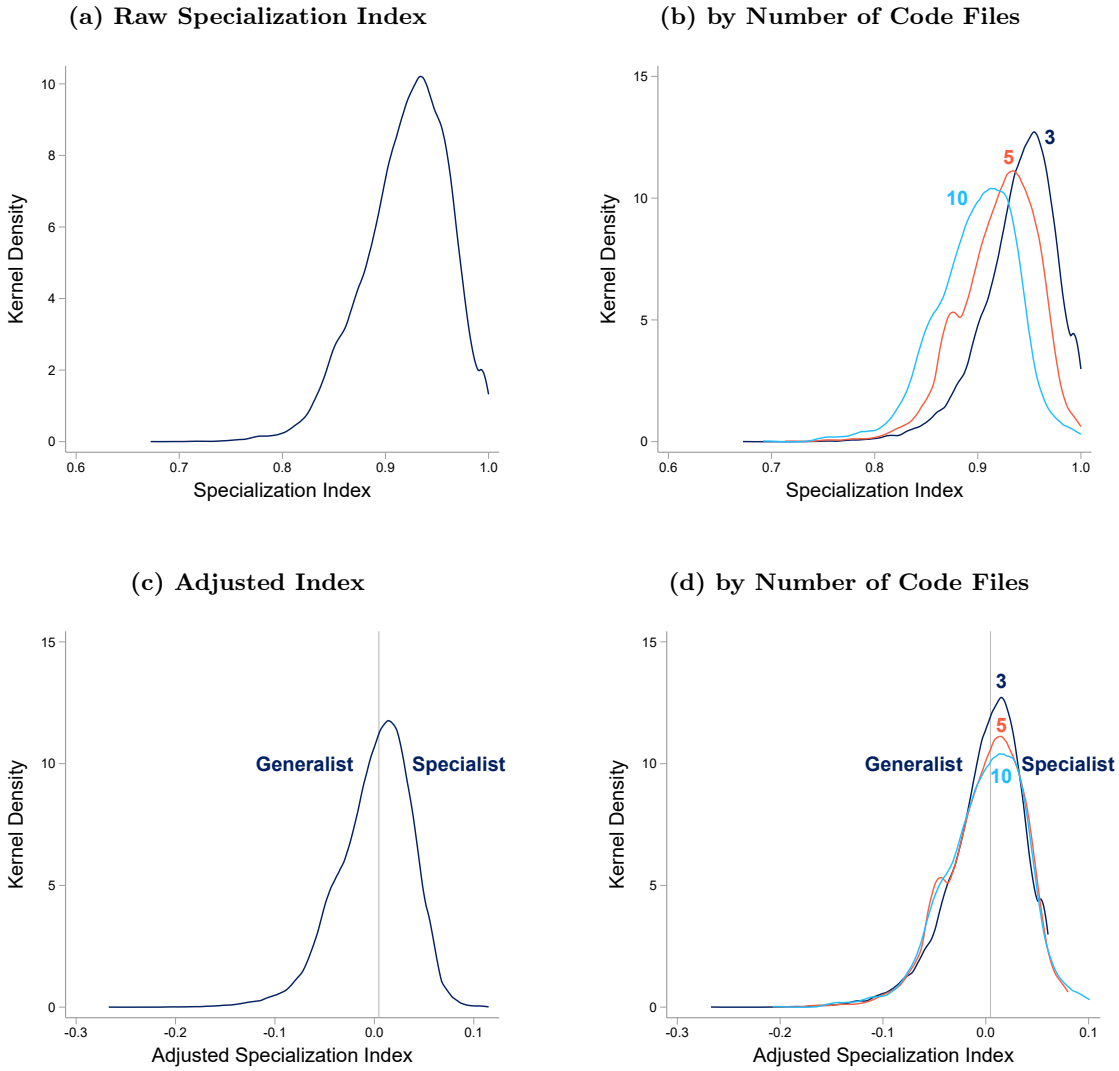
Quality I measure solution quality in three different ways, natural to the setting: *Rank*, *Medal Win*, and *Top 3*. These three metrics roughly target three different objects: average quality, high quality, and top quality. First, all competitions on Kaggle specify an objective, numerical target to assess a submission's quality, for example, minimizing root mean square error. All submissions are then ranked according to this metric. This allows me to measure quality differences in a continuous, granular manner. *Rank* transforms a team's absolute rank in a competition into their percentile rank within that competition. *Rank* is equal to 100 for the best, and 0 for the worst performing team. This allows me to compare teams' relative positions across competitions with different numbers of competitors. Second, a share of high-ranking teams receive virtual gold, silver, or bronze medals. I generate a binary indicator to measure differences in the probability of attaining a medal, *Medal Win*.[17] Third, I create an indicator for placing among the top three competitors. The top three competitors are typically awarded monetary prizes, required to license their solution code, and publish detailed, often paper-like solutions on Kaggle. We can think of winning a medal as providing a high-quality solution, and placing among the top three as an excellent solution.

[15] 768 tokens ("words" of code) to be exact.

[16] If these sub-units are still too large to be processed by Code T5+, I split them further into lines.

[17] Kaggle users refer to this as placing "in the metal". The number of medals awarded in a competition is determined by the number of competitors. For details see Kaggle.com (2025).

Figure 1.1: Specialization Index



Notes: The figure illustrates the construction of the specialization index. Panel (a) displays the distribution of the raw specialization index, which is heavily skewed towards high similarities. Panel (b) illustrates how the number of code files affects the distribution of the raw specialization index with the examples of users who have written 3, 5, or 10 code files. More code files mechanically lead to lower specialization. Panel (c) shows the distribution of the number-of-code-files adjusted specialization, removing the impact of the number of code files on the distribution of specialization (Panel (d)).

Complexity I assess a competition's (problem's) complexity by measuring the ex-post likelihood of overfitting. To uncover complex relationships, a machine learning model needs to be able to incorporate this complexity. The higher a model's capacity to reflect complex relationships, the higher the complexity of the model itself. However, model complexity comes at the cost of a higher risk of overfitting (Mohri et al., 2018). The algorithm might pick up spurious correlations in the

test data, but fail to capture true underlying relationships (Mullainathan and Spiess, 2017). A competition is doubly complex if it is hard to spot this risk of overfitting. I measure a competition's complexity as the aggregate realized amount of overfitting. To measure overfitting risk, I use "Shake-Up", a concept coined by Kaggle users to assess overfitting in competitions (Trotman, 2019), which has been validated as a measure of overfitting in machine learning (Roelofs et al., 2019):

$$\text{Complexity} = \frac{1}{\text{Number of Competitors}} \sum \frac{|\text{Private Rank} - \text{Public Rank}|}{\text{Number of Competitors}} \tag{1.7}$$

In other words, the average change in relative performance of the algorithm from public to private test set. Large changes are mostly driven by many competitors overfitting on the training data and achieving high scores on the public dataset, but dropping in ranks once the algorithm is evaluated against the test data. Those that did not overfit will subsequently shoot up in ranks, even if the score of their algorithm remains stable – as it should when properly designed.

1.3.4.2 Additional Variables

User Characteristics To control for factors that may influence solution quality other than specialization, I define a range of time-varying user-level covariates. To proxy a user's current ability, I generate two measures of performance: *Previous Rank* and *Previous Medal Wins*. *Previous Rank* is the average percentile rank achieved by a user in prior competitions in which they participated as a solo competitor. *Previous Medal Wins* is the share of previous competitions in which a user won a medal among all competitions in which they participated as a solo competitor. I also measure *Competition Experience*, the number of competitions a user competed in, whether alone or in a team, before entering a given competition. I collect an additional set of user demographics from users' profiles, on which users can report their location, occupation, and preferred pronouns, as well as write short biographical notes (*user bio*).[18] I use pronouns and user names to classify a user's gender in a multi-step procedure, for details, see Appendix 1.B.2. I hand-code occupation strings and occupational information reported in the user's bio into 17 occupational categories. Appendix Table 1.B.2 provides details on chosen occupational categories and included occupations.

Code Characteristics Since specialization is measured from code, I include several characteristics of code files: The number of lines per code file, the number of lines of comment, and the votes received by a code file as a measure of code quality.[19] I aggregate code characteristics at the author level.

[18] Many users report CV-like information in their bio, some share their motivation to participate in Kaggle, others only include pictures or emojis.

[19] If another user finds a code file useful, well-written, or creative, they can vote for the code file. Since Kaggle status groups are based on votes, voting behavior is strictly monitored. Self-promotion or vote collusion can result in bans (see Kaggle.com (2024b)), so votes are a relatively noise-free signal of code quality.

Team Characteristics Team characteristics are simply all team members' user-level characteristics aggregated to the team level as well as team size. This mirrors the identification strategy in Weidmann and Deming (2021), where individual performance at the same task of all team members is aggregated to isolate the effect of team members' ability from team effects.

Submission Characteristics During a competition, participants can submit preliminary solutions to the competition, which are then evaluated against the public test set. For all submissions, I observe which user made the submission, submission dates, algorithm performance, and which submissions were chosen for the final submission. I use submission data in two ways: First, since I am able to construct counterfactual ranks for each preliminary submission – information that is not available to competition participants – I investigate the change in submission quality in response to teaming up in an event study approach similar to Lemus and Marshall (2024). Second, I use the amount and timing of submissions as a measure for motivation and effort to rule out alternative mechanisms.

1.3.5 Sample Construction and Summary Statistics

I start with the 7.19 million team member observations who participate in any competition. Note that teams are defined at the competition level, i.e., team member refers to a user on a specific team in a specific competition. I then drop teams who participate in competitions in which performance is not incentivized, as well as recruitment competitions, in which there naturally is no teamwork. This leaves me with 4.35 million observations. I then drop teams that sign up for a competition but never make a submission, which is the majority of teams (3.83 million). A possible explanation for this strikingly large number is that, to access competition data, Kaggle users have to formally sign up for a competition, even if they never intended to compete.[20] I also drop all benchmark submissions by competition hosts and the Kaggle team. At the end, 518,240 team member observations remain. In a final step, I drop all teams for which I cannot measure specialization for one or more team members, to ensure I accurately measure team types. This leaves me with 76,776 user-competition observations, or 74,433 teams, from 357 distinct competitions. These are comprised of 13,646 distinct users. Table 1.1 presents summary statistics for competitions.

Table 1.2 presents summary statistics for users, split into generalists and specialists. Kaggle users are mostly male, have been active on the platform for about 2.5 years, and enter many competitions, but compete mostly on their own. Observing many solo competitions per user allows me to measure individual ability independent of team performance. Generalists have been active on the platform slightly longer than specialists, are more likely to report their location and to be from

[20] Of these, 49% are teams that sign up for a competition after the competition deadline, which is possible since Kaggle competitions stay active after they ended to enable other users to access the competition data.

Table 1.1: Summary Statistics on Competitions

Variable	Mean	SD	Observations
Prizes	53152.26	140325.19	353
Competitors	1450.75	1493.91	357
Competitors with Specialization	215.06	223.78	357
Duration	86.49	48.85	357
Teams	0.02	0.02	357
Research	0.34		357
Post ChatGPT	0.11		357

Notes: The table reports summary statistics for competitions. Data come from Meta Kaggle (Risdal and Bozsolik, 2022). Prizes are total competition prize pools in U.S. dollars. Competitors reflect the total number of competitors submitting to a competition, and Competitors with Specialization are the subset of competitors for which I can measure specialization. Duration is the duration of a competition in days. Teams indicates what share of competitors enters the competition as a team. Research is an indicator whether the competition is a research competition (as defined by Kaggle). Post ChatGPT is an indicator whether a competition ended after the introduction of ChatGPT.

Table 1.2: Summary Statistics on Users

Variable	All Users	Generalists	Specialists	p-value Gen = Spec
No. Users	13646	7077	6569	
Female	0.068	0.065	0.071	0.18
User Platform Age	940.001	1006.982	868.054	0.00
Reports Location	0.856	0.876	0.835	0.00
India	0.220	0.210	0.231	0.00
USA	0.134	0.161	0.105	0.00
Reports Occupation	0.680	0.692	0.667	0.00
Data Scientist	0.164	0.175	0.151	0.00
Code Files	10.628	10.318	10.962	0.25
Lines	219.740	188.315	253.595	0.00
Votes	6.308	6.572	6.025	0.15
Competitions	5.626	6.143	5.070	0.00
Collaborations	0.048	0.046	0.052	0.05
Mixed Team	0.571	0.547	0.600	0.02
Team Lead	0.423	0.433	0.412	0.31

Notes: The table reports summary statistics for users. Data come from Meta Kaggle (Risdal and Bozsolik, 2022) and own hand-coding. Female is an indicator whether the user is female, user platform age is the time (in days) a user has been active on Kaggle. Reports location is an indicator whether the user reports a location in their profile, with India (dummy) and USA (dummy) being the two most common locations. Reports Occupation is an indicator whether the user includes information on their occupation in their profile, with Data Scientist (dummy) being the most frequent occupation. Code Files is the number of code files solo-written by a user, lines the average length of a code file, and votes the average number of votes received by a code file. Competitions is the number of competitions a user in the data enters, whether as a solo competitor or in a team. Collaborations indicates the share of competitions a user enters as a team. Mixed Team measures the share of a user's teams that are mixed teams, and team lead is the share of teams in which a user is the team lead.

the USA, to report their occupation and to identify as data scientists, write shorter code, are less likely to collaborate, and to do so in mixed teams. The empirical analysis take these underlying differences into account.

Table 1.3: Summary Statistics on Stable Teams

Variable	All Teams	Solo Competitor	Generalist Team	Mixed Team	Specialist Team
No. Team-Competition Observations	74433	72721	482	922	308
No. Stable Collaborations	14568	13326	327	704	211
No. Competitions as Collaboration	5.11	5.46	1.47	1.31	1.46
Rank	46.45	43.75	77.21	76.83	67.86
Medal Win	0.11	0.07	0.49	0.48	0.34
Top 3	0.00	0.00	0.06	0.04	0.02
Team Size	1.12	1.00	2.21	2.63	2.21
Female Share	0.06	0.07	0.05	0.05	0.06
Data Scientist	0.17	0.16	0.28	0.24	0.22
At 1st Competition as Collaboration					
User Platform Age	504.88	467.77	995.60	902.37	735.08
Experience	1.23	0.19	13.21	13.19	8.44
Previous Individual Rank	61.20		62.58	61.73	56.69
Previous Medal Wins	0.22		0.24	0.22	0.20
Total Submissions	18.46	12.68	77.73	85.60	67.54
High Complexity Competition	0.29	0.30	0.28	0.28	0.29
Post ChatGPT Competition	0.17	0.18	0.11	0.16	0.17

Notes: The table reports summary statistics for stable teams, i.e., teams with the exact same members across competitions. One user can be a member of multiple stable teams, and a solo competitor is also considered a stable team whenever they compete alone. Data come from Meta Kaggle (Risdal and Bozsolik, 2022). No. Competitions as Stable Team measures how frequently I observe a team composed of the exact same users across competitions. Rank is the percentile rank achieved by a team in a competition, Medal Win an indicator for whether a team won a medal, and Top 3 an indicator for whether a team placed among the top 3 in a competition. Team Size measures the size of a team, female share is the share of team members who are female, and Data Scientist the share of team members whose occupation is Data Scientist. User Platform Age is the team member's average time active on the platform when they first enter a competition (for solo competitors) or first join a team. Experience measures the team member's average number of previous competitions. For solo competitors, this is the number of competitions participated in as a team before the first solo competition, for teams this is the average number of competitions participated in as a solo competitor or in a team before joining the stable team. By construction, Previous Individual Rank and Previous Medal Win are not defined for a solo competitor's first solo competition. For teams, these reflect the average percentile rank achieved by team members in solo competitions, and the team-member-average of Previous Medal Wins, i.e., the share of previous solo competitions in which a team member has won a medal. Total Submissions is the count of submissions made to a competition by a team. High Complexity Competitions is an indicator that is one whenever a team enters a high complexity competition (for details, see Section 1.3.4.1). Post ChatGPT is an indicator that is one whenever a team enters a competition that ended place after the introduction of ChatGPT.

Table 1.3 presents summary statistics for stable teams. I define a team composed of the exact same users across competitions as a stable team. Note that I also consider a user's observations from solo competitions as stemming from a stable "team". Overall, I observe 327 generalist teams who compete 1.47 times on average, 704 mixed teams competing 1.31 times on average, and 211 specialist teams competing on average 1.46 times. Teams on Kaggle are small. While most competition entrants are solo competitors, teams are mostly comprised of two people. Users who compete in teams have been active on the platform for a longer time, and users tend to enter many competitions

on their own before ever competing as a team. Teams also make more submissions than solo competitors and outperform solo competitors on all dimensions. Among teams, generalist teams achieve the highest ranks, win most medals, and place among the top 3 most often. Reflecting the higher experience of generalists, generalist teams also are comprised of the most experienced and able users. Isolating the effect of the ingredients – team members – and the recipe – team type – on algorithm quality is the focus of the next section.

1.4 The Impact of Team Type on Solution Quality

1.4.1 Empirical Strategy

The quality of a team's algorithm naturally depends both on the characteristics of individual team members as well as any synergies created by the team. Therefore, it is crucial to isolate the effects of individual team members from the overall team effect. For instance, if generalists tend to be more experienced, the superior performance of generalist teams might reflect experience rather than any inherent benefit of the team structure itself. I employ several strategies to disentangle the effect of individual team members from the composition effect. This is econometrically challenging, as outcomes are realized at the team level, but potential confounds arise at the individual level (Constantine and Correia, 2021; Bonhomme, 2022).

A particular concern is the ability of individual team members, for which I first try to account explicitly by constructing a measure of individual ability from team members' past solo competition performance. Adhvaryu et al. (2023) use a similar approach to isolate the effect of individual scientists on co-authored papers or joint patents. To address unobservable team member heterogeneity, I include a specification with team member fixed effects. This is similar in spirit to the AKM literature (Abowd et al., 1999), where worker fixed effects are identified from firm switchers. I identify scientist fixed effects from team switchers as well as from switches from solo work to teamwork. Finally, I utilize the fact that I can observe individual contributions to team output in the form of submissions. For a small subset of teams, I can track individual team member's submissions over the course of the competition and assess the impact of joining a team of a specific type on submission quality, a strategy similar to the one implemented by Lemus and Marshall (2024).

1.4.2 Results Isolating Team Member Ability

I first estimate a regression of solution quality on team type to test Prediction 1:

$$\mathrm{Y}_{tc} = \alpha + \tau_1 \text{Generalist Team}_t + \tau_2 \text{Mixed Team}_t + \tau_3 \text{Specialist Team}_t + X'_{tc}\beta + \gamma_c + \epsilon_{tc} \qquad (1.8)$$

Where Y_{tc} is a team t's solution quality (Rank, Medal Win, Top 3) in competition c, α is the intercept, equivalent to the average performance of a solo competitor. Generalist Team$_t$ is an indicator that equals one if team t is comprised of only generalists, Mixed Team$_t$ is an indicator that equals one if team t consists of both generalists and specialists, and Specialist Team$_t$ is an indicator that equals one if team t consists of only specialists. The coefficients τ_1, τ_2, and τ_3 measure the difference in solution quality of each type of team relative to the solution quality of a solo competitor. X_{tc} is a vector of time-varying team-level controls including average team member experience, proxies for average team member abilities, team size, and demographic controls, including occupation shares for all occupations reported by at least 100 users, country shares for all countries with at least 100 users, share female, and the average time team members have been active on the platform. γ_c are competition fixed effects.

Table 1.4 displays results for Equation (1.8) on the percentile rank achieved by each team's solution. Column 1 shows gross quality differences: all teams produce higher quality than individual competitors, with generalist teams placing 21 percentile ranks higher than solo competitors, mixed teams placing roughly 18 percentile ranks above solo competitors, and specialist teams placing about 12.5 percentile ranks above solo competitors. These differences are significant at the 1% level. In addition, the quality differences between team types are also highly significant, with generalists achieving higher quality than mixed and specialist teams.

As previously discussed, quality differences might simply be a reflection of pre-existing ability differences and positive selection into collaborating rather than any effect of teamwork or the specific type of team. To account for these differences, column 2 adds controls for team members' previous competition experience and previous percentile ranks achieved by team members when they were competing on their own. This reduces sample size by circa 14,000 observations, since I can only observe experience and previous ranks for teams that have at least one competitor that has competed once before. While the effect of competition experience is significant but small, the effect of previous rank is large in size: Placing 1 percentile higher in previous competitions is associated with placing 0.5 percentiles higher in the current competition's ranking on average. Again, teams consistently achieve higher quality than solo competitors. The differences between team types however are smaller, reducing performance gaps to 3 percentile ranks between generalist and specialist teams, and 2 percentile ranks between generalist and mixed teams. This reflects the fact that generalist teams seem to be positively selected (c.f. Table 1.2 and Table 1.3). In column 3, I add the full set of demographic controls with minimal effects on coefficient sizes. Since specialization is measured from code, my index might also capture subtle differences in code quality, which could be correlated with team performance. Column 4 thus adds code quality controls, such as the length of code, amount of comments, the number of code files, and votes received by code

Table 1.4: Team Type and Solution Quality: Rank

Dependent Variable:	*Rank*				
	(1)	(2)	(3)	(4)	(5)
Generalist Team	21.19***	18.45***	18.77***	18.66***	20.03***
	(1.44)	(1.11)	(1.08)	(1.07)	(1.07)
Mixed Team	17.85***	16.16***	16.46***	16.20***	16.65***
	(1.12)	(0.94)	(0.93)	(0.92)	(0.90)
Specialist Team	12.50***	15.73***	16.26***	15.92***	15.80***
	(2.09)	(1.99)	(2.02)	(2.04)	(1.99)
Competition Experience		0.03***	0.05***	0.05***	0.08***
		(0.01)	(0.01)	(0.01)	(0.01)
Previous Rank		0.52***	0.50***	0.49***	0.49***
		(0.01)	(0.01)	(0.01)	(0.01)
3-Person Team	8.11***	6.71***	6.38***	6.11***	6.07***
	(2.00)	(1.63)	(1.62)	(1.60)	(1.58)
4-Person Team	17.18***	12.98***	12.49***	12.30***	12.47***
	(2.26)	(2.34)	(2.35)	(2.31)	(2.19)
5-Person Team	17.15***	13.74***	13.58***	13.48***	14.69***
	(2.13)	(1.69)	(1.69)	(1.67)	(1.74)
R^2	0.01	0.15	0.16	0.16	0.20
Observations	74433	60897	60204	60090	60090
Demographic Controls			Yes	Yes	Yes
Code Quality Controls				Yes	Yes
Competition FEs					Yes
p-value (Generalist Team = Mixed Team)	0.04	0.09	0.08	0.06	0.01
p-value (Generalist Team = Specialist Team)	0.00	0.22	0.26	0.22	0.06
p-value (Specialist Team = Mixed Team)	0.02	0.84	0.92	0.89	0.68

Notes: The table reports the estimates of equation (1.8). The dependent variable measures the percentile rank of a team's solution in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Competition experience measures the average of the number of competitions each team member participated in before competition c. Previous Rank measures the average of the percentile rank achieved by each team member when competing alone in a previous competition. Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

files as a proxy for code quality. The results are quantitatively similar to columns 2 and 3.

In the model, a crucial determinant of quality gaps between generalist and specialist teams is a problem's type. Thus, I add competition fixed effects in column 5. When accounting for problem type, the quality difference between generalist teams and specialist teams are large and significant: generalist teams place around 4 percentile ranks higher than specialist teams, this difference is significant at the 10% level. Generalist teams also place about 3.5 percentile ranks higher than mixed teams (significant at the 1% level).

Are generalist teams also more likely to provide high quality solutions? Table 1.5 displays results for the probability of winning a medal. As in Table 1.4, column 1, all team types are significantly more likely to win a medal than solo competitors. Generalist teams are 31 percentage points more likely to win a medal, while mixed and specialist teams are 25 and 15 percentage points more likely. These coefficients are all significantly different from each other (at the 5 or 1 % level). When controlling for experience and previous medal wins in column 2, the coefficient for

Table 1.5: Team Type and Solution Quality: Medal Win

Dependent Variable:	*Medal Win*				
	(1)	(2)	(3)	(4)	(5)
Generalist Team	0.31***	0.29***	0.30***	0.30***	0.30***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Mixed Team	0.25***	0.24***	0.25***	0.25***	0.25***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Specialist Team	0.15***	0.20***	0.22***	0.21***	0.22***
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Competition Experience		-0.00**	-0.00***	-0.00***	-0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Previous Medal Wins		0.48***	0.45***	0.43***	0.39***
		(0.01)	(0.01)	(0.01)	(0.01)
3-Person Team	0.19***	0.18***	0.18***	0.17***	0.18***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
4-Person Team	0.32***	0.29***	0.28***	0.28***	0.30***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
5-Person Team	0.29***	0.28***	0.27***	0.27***	0.29***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
R^2	0.02	0.11	0.12	0.12	0.15
Observations	74433	60897	60204	60090	60090
Demographic Controls			Yes	Yes	Yes
Code Quality Controls				Yes	Yes
Competition FEs					Yes
p-value (Generalist Team = Mixed Team)	0.07	0.10	0.12	0.11	0.08
p-value (Generalist Team = Specialist Team)	0.00	0.03	0.05	0.04	0.04
p-value (Specialist Team = Mixed Team)	0.03	0.28	0.37	0.36	0.37

Notes: The table reports the estimates of equation (1.8). The dependent variable is an indicator whether a team's solution has won a medal in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Competition experience measures the average of the number of competitions each team member participated in before competition c. Previous Medal Wins measures the average of the share of previous competitions for each team member in which a team member won a medal when competing alone. Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

specialist teams increases to 20 percentage points. However, it remains significantly lower than the effect for generalist teams at 29 percentage points. Competitors' medal track records are highly predictive of winning a medal: If all team members have won a medal before, their probability of winning a medal as a team is 48 percentage points higher than that of a team in which no member has ever won a medal. Controlling for demographics (column 3), code quality (column 4), and including competition fixed effects (column 5) barely affects coefficient sizes. When accounting for all covariates and fixed effects, generalist teams are 8 percentage points more likely to win a medal than specialist teams.

Turning now to the probability of providing an excellent solution, Table 1.6 reports results for Equation (1.8) on the likelihood of a team's solution being among the top three solutions in a competition. Column 1 displays gross quality differences. While generalist teams are 5 percentage points more likely to find a solution that is among the top three, and mixed teams are 1 percentage

Table 1.6: Team Type and Solution Quality: Top 3

Dependent Variable:	*Top 3*				
	(1)	(2)	(3)	(4)	(5)
Generalist Team	0.05***	0.06***	0.06***	0.06***	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Mixed Team	0.01***	0.02**	0.02***	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Specialist Team	0.00	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Competition Experience		-0.00*	-0.00***	-0.00***	-0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank		0.00***	0.00***	0.00***	0.00***
		(0.00)	(0.00)	(0.00)	(0.00)
3-Person Team	0.04**	0.04**	0.04**	0.04**	0.04**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
4-Person Team	0.05*	0.05	0.05	0.05	0.05
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
5-Person Team	-0.00	-0.01	-0.01	-0.01	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
R^2	0.01	0.02	0.02	0.03	0.05
Observations	74433	60897	60204	60090	60090
Demographic Controls			Yes	Yes	Yes
Code Quality Controls				Yes	Yes
Competition FEs					Yes
p-value (Generalist Team = Mixed Team)	0.00	0.00	0.00	0.00	0.00
p-value (Generalist Team = Specialist Team)	0.00	0.00	0.00	0.00	0.00
p-value (Specialist Team = Mixed Team)	0.24	0.50	0.54	0.52	0.52

Notes: The table reports the estimates of equation (1.8). The dependent variable is an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Competition experience measures the average of the number of competitions each team member participated in before competition c. Previous Rank measures the average of the percentile ranks achieved by each team member when competing alone in a previous competition. Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

point more likely, specialist teams are indistinguishable from solo competitors. Controlling for experience and previous performance (column 2), demographics (column 3), code quality (column 4), and including competition fixed effects (column 5) has minimal effect on coefficient sizes and significance. In fact, generalist and mixed teams' advantage over solo competitors increases by one percentage point compared to column 1, while specialist teams are not differentially likely to place among the top three than solo competitors. Generalist teams are significantly more likely to place among the top three than mixed teams and specialist teams throughout all specifications. These differences are significant at the 1% level. Taken together, the patterns documented in Tables 1.4-1.6 provide strong support for Prediction 1.

1.4.3 Results Isolating Team Member Unobservables

A potential concern is that previous solution quality achieved by team members as solo competitors and code quality do not sufficiently capture individual ability. Rather, there may be time-invariant unobservable user characteristics that influence team solution quality. To address this concern, I estimate the following regression:

$$\mathrm{Y}_{tc} = \alpha + \tau_1 \text{Generalist Team}_t + \tau_2 \text{Mixed Team}_t + \tau_3 \text{Specialist Team}_t + \gamma_c + \sum_{i \in G(t)} \upsilon_i + \epsilon_{tc} \quad (1.9)$$

Y_{tc} and Generalist Team_t, Mixed Team_t, and Specialist Team_t are defined as above. $\sum_{i \in G(t)} \upsilon_i$ is a function which aggregates all $G(t)$ user fixed effects υ_i to the team level, γ_c are competition fixed effects, and ϵ_{tc} are team-competition specific shocks. To efficiently estimate individual fixed effects with group level outcomes, I use the estimator introduced with the reghdfe package (Constantine and Correia, 2021), which aggregates individual fixed effects at the group level and provides accurate inference.

Table 1.7: Team Type and Solution Quality: Team Member Fixed Effects

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	17.98***	19.25***	0.28***	0.28***	0.05***	0.05***
	(1.00)	(1.05)	(0.02)	(0.02)	(0.01)	(0.01)
Mixed Team	17.72***	18.17***	0.26***	0.27***	0.02***	0.02***
	(0.90)	(0.88)	(0.02)	(0.02)	(0.01)	(0.01)
Specialist Team	16.23***	15.82***	0.22***	0.23***	0.02	0.02
	(1.56)	(1.51)	(0.03)	(0.03)	(0.01)	(0.01)
3-Person Team	5.05***	5.14***	0.16***	0.16***	0.04**	0.04**
	(1.32)	(1.31)	(0.03)	(0.03)	(0.02)	(0.02)
4-Person Team	9.43***	8.99***	0.26***	0.27***	0.04	0.04
	(1.89)	(1.93)	(0.05)	(0.05)	(0.03)	(0.03)
5-Person Team	10.06***	10.97***	0.28***	0.29***	-0.00	-0.00
	(1.74)	(1.87)	(0.06)	(0.06)	(0.02)	(0.02)
R^2	0.36	0.40	0.26	0.29	0.14	0.16
Observations	69381	69372	69381	69372	69381	69372
User FEs	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs		Yes		Yes		Yes
p-value (Generalist Team = Mixed Team)	0.84	0.39	0.56	0.51	0.05	0.04
p-value (Generalist Team = Specialist Team)	0.34	0.06	0.17	0.21	0.04	0.04
p-value (Specialist Team = Mixed Team)	0.40	0.17	0.30	0.40	0.69	0.70

Notes: The table reports the estimates of equation (1.9). The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.7 displays results for Equation (1.9). Columns 1 and 2 report coefficients for the team's percentile ranks, once with (1) and once without (2) including competition fixed effects. The

coefficients on the different team types are close to the coefficients reported in Table 1.4. This pattern is repeated in columns 3 and 4, which report results for the probability of a team's solution winning a medal, and columns 5 and 6, which report results for the probability of a team's solution placing among the top three solutions in a competition. In all specifications, generalist teams provide higher quality solutions than specialist teams. The difference between generalist and specialist teams is significant at the 10% level for Rank, and at the 5% level for Top 3, but not for Medal Win, despite similar coefficient sizes as in Table 1.5. These results indicate that quality differences between generalist and specialist teams are not driven by time-invariant unobservable user characteristics.

1.4.4 Results Isolating the Effect of Joining a Team

In the previous sections, I have shown that generalist teams achieve higher quality solutions than specialist teams, and that this difference is driven neither by time-varying observables nor time-constant unobservables. Both empirical approaches relied exclusively on between-competition variation. Still, one might be concerned that performance in different competitions insufficiently captures task-specific ability. Using performance in other competitions as a proxy for ability in a given competition implicitly assumes that computer scientists only choose to enter competitions that match their skill sets. If, for instance, computer scientists are not fully aware of which skills a problem requires before entering the competition, generalists would in expectation work on problems that are within their skill set more often, since their skill set is broader. This is not an unlikely scenario. Furthermore, generalists' broader skills might enable them to better recognize which skills are required by a competition.

To address this concern, I make use of two unique features on Kaggle, preliminary submissions and in-competition team formation. As described in Section 1.3.4.2, competitors are able to make preliminary submissions to a competition to receive feedback on how the proposed algorithm performs on the public test set. While only performance on the public, but not the private (final) test set is known to competitors during the competition, Kaggle also calculates and records how an algorithm would have scored on the final test set. Using these scores and information on the scoring algorithm, I can construct counterfactual percentile ranks for each submission.

Users do not have to enter a competition as a team, but can also first enter a competition as a solo competitor, work independently, and later form a team. Importantly, once they join a team, which team member made a submission is still recorded by Kaggle. Similarly to Lemus and Marshall (2024), I track how submission quality changes before and after team formation. However, I deviate from their analysis in two core dimensions. First, my analysis is at the team *member* level rather than at the team level, allowing me to isolate individual contributions to team performance.

Second, since the timing of team formation is likely endogenous, I restrict my analysis to teams forming at the exogenously set team formation deadline.[21] The team formation deadline is the modal date of team formation for all team types, see Figure 1.C.2. Since team members can join a team in a sequential way, "treatment" timing in larger teams is ambiguous, such that I additionally drop all teams with more than two members.
I then estimate the following event study:

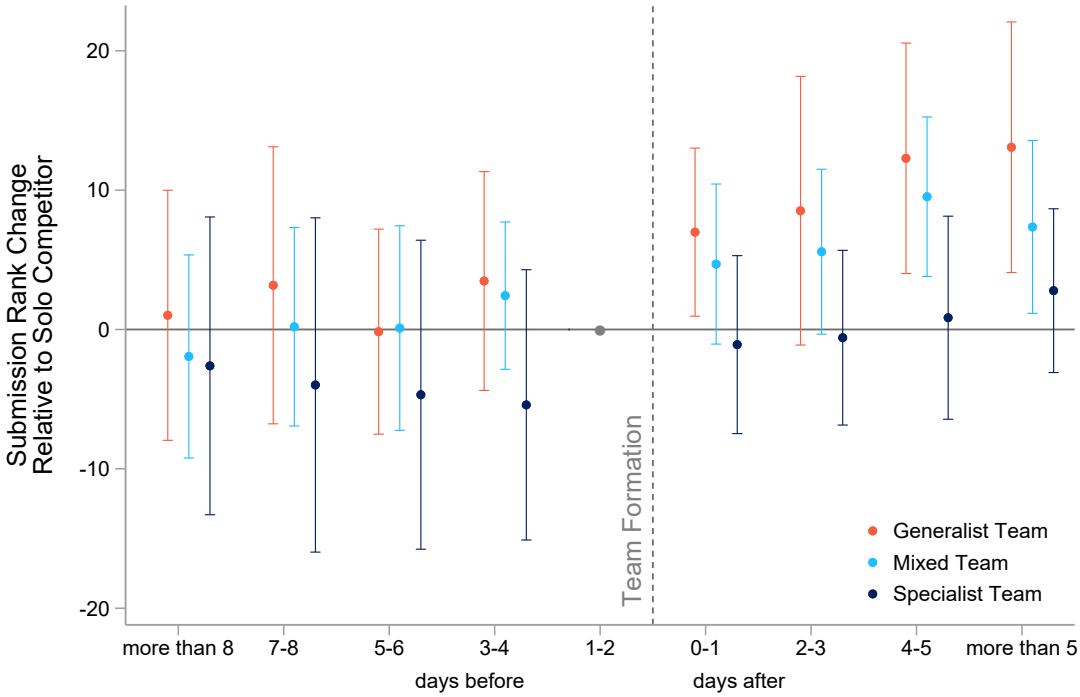
$$\begin{aligned}\text{Submission Rank}_{ict} =&\text{Generalist Team}_{ic} \sum_{d\neq -1} \delta_d^G D_{T+d} + \text{Mixed Team}_{ic} \sum_{d\neq -1} \delta_d^M D_{T+d} \\ &+\text{Specialist Team}_{ic} \sum_{d\neq -1} \delta_d^S D_{T+d} + \sum_{d\neq -1} \delta_d D_{T+d} + \upsilon_{ic} + \phi_t + \epsilon_{ict} \end{aligned} \tag{1.10}$$

where d indexes periods of two days relative to team formation, and D_{T+d} denote the event-study indicators for the periods leading up to and following team formation. Generalist Team$_{ic}$ is an indicator equal to one if user i joins a generalist team, Mixed Team$_{ic}$ is an indicator equal to one if user i joins a mixed team, and Specialist Team$_{ic}$ is an indicator equal to one if user i joins a specialist team team in competition c. Coefficients δ_d capture the effect of time relative to the team formation deadline at T for users who never form a team. δ_d^G measure the differential effect of time relative team formation for generalist teams, δ_d^M for mixed teams, and δ_d^S for specialist teams. υ_{ic} are team member fixed effects, i.e., a user $\times$ competition fixed effect, and ϕ_t are competition day fixed effects.

Figure 1.2 plots results from Equation (1.10), and Table 1.C.3 displays coefficients. Since the team formation deadline most commonly coincides with the week before the competition ends, I include three lags of two days, and one lag pooling all submission dates more than five days after the team formation deadline. As Figure 1.C.3 illustrates, few teams make submissions after day seven. I also include two-day leads only for up to eight days before the team formation deadline and one lead capturing all earlier submission dates, since submissions are comparatively sparse in this time. The small number of daily submissions also informs my choice to bin submission dates into two-day windows, since estimation of the multitude of fixed effects is demanding. Before forming a team, submission ranks of eventual team members develop no differently than those of solo competitors, indicating no anticipation. After forming a team, the quality of submissions made by the members of generalist teams increases immediately by about 7 percentile ranks relative to users who do not form a team. This effect is significant at the 5% level, and significantly different from mixed and specialist teams, for whom I cannot detect any effects of team formation on users working in mixed or specialist teams. In the week after joining a team, submission quality by team

[21] This specification choice also allows me to circumvent issues created by staggered treatment as discussed in Sun and Abraham (2021), as all competitors in my analysis either form a team at exactly the same time or not at all.

Figure 1.2: Team Type and Solution Quality: Submission Event Study



Notes: The figure plots the estimated coefficients δ_d^G, δ_d^M, and δ_d^S from Equation (1.10). Each coefficient is equivalent to a two-day period d before or after team formation at the exogenously set team formation deadline. The two-day period immediately before team formation is the omitted category. The coefficients indicate the percentile difference in submission ranks for each team type relative to the team formation deadline and to solo competitors who do not form a team at $d = 0$. The corresponding regression results are reported in Table 1.C.3.

members in generalist teams steadily increases relative to solo competitors, with members of mixed teams catching up slightly. The quality of submissions by members in specialist teams does not change significantly in response to team formation relative to those by solo competitors. After five days, submissions by team members of generalist teams place circa 10 percentile ranks higher than those of members of specialist teams.

1.4.5 Robustness

Robustness to Ability Proxies In Section 1.4.2 I use the average of past performance in all solo competitions to control for potential ability differences between team types. This could potentially over- or understate current ability if computer scientists' skills change dynamically, for instance because they learn from past competitions. To address this concern, I include only recent performance as an ability control in Table 1.C.4. Specifically, I average the percentile ranks achieved

in the last (respectively second to last and third to last) solo competition for all team members. For medals, I take the share of team members who achieved a medal in their last (second/third to last) competition. We might also be worried that rather than the absolute past performance, trajectories matter for team performance. Table 1.C.5 controls for the team-level average change in percentile ranks between the first and second to last, second and third to last, and the average of first and second and second and third to last solo competitions. Coefficient sizes remain remarkably stable across specifications, as does significance. The only exception are results for medal win with three performance lags (Table 1.C.4, Column 8) where I cannot detect significant differences between generalist and specialist teams anymore, likely due to substantially reduced sample size.

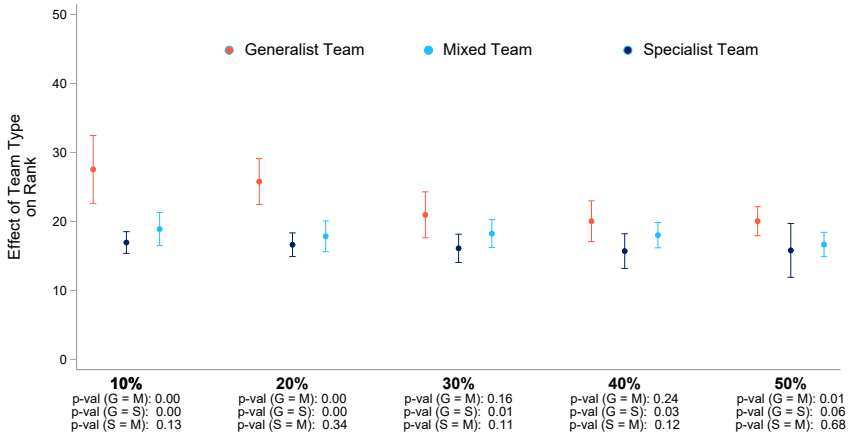
Robustness to Specialization Measure To ensure results are not driven by how I chose to classify specialists and generalists, I conduct several robustness checks. First, because I create generalist and specialist categories by binarizing the continuous specialization index, I conduct a sensitivity analysis to determine how different binarization cutoff points affect my results. Figure 1.3 presents results of this sensitivity analysis. It plots coefficients from Equation (1.8), including all controls and competition fixed effects as in Column (5) of Tables 1.4-1.6 with alternative team type classifications. Each category on the x-axis corresponds to a different cutoff point in the specialization index, e.g., 10% reflects team type classifications in which only individuals with very generalist skills, i.e., only those whose specialization is in the lowest decile of the specialization index were classified as generalists. Results at 50% correspond to the main results presented in this section. For all cutoff points below 50%, generalist teams achieve significantly higher ranks (Figure 1.3a), are more likely to win a medal (Figure 1.3b), and to place among the top three (Figure 1.3c) than specialist teams. In fact, the stricter I classify generalists, the larger the advantage of generalist teams over specialist and mixed teams, suggesting the results presented in this section might be a conservative estimate of the differences between generalists and specialists.

As a further robustness check, I re-calculate the specialization index excluding all code files published after November 30$^{\text{th}}$ 2022, when ChatGPT was introduced. This reduces the sample size by about 4000 teams. I also construct an alternative measure of specialization based on the python packages included in each code file. To use the functionality of a given python package, each python code file needs to explicitly import a package at the beginning of the file. Imported packages correspond to the goal of the code, for instance, package cv2 provides functionality for computer vision problems such as image processing, whereas package nltk is used for natural language processing. I construct *Python Package Diversity* (PPD) as:

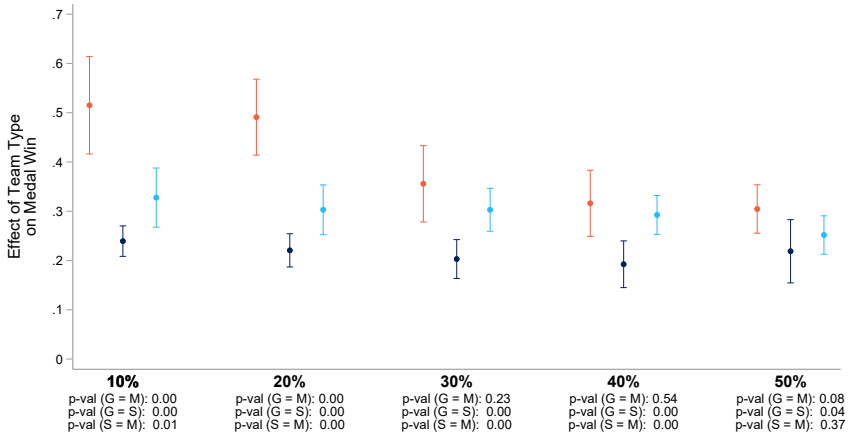
$$PPD_i = \frac{\sum_{p=1}^{P_i} \text{Share of user i's Code Files using Package p}}{\text{Number of Distinct Packages used by user i}}$$

Figure 1.3: Robustness of Main Result to Different Classification Cutoffs

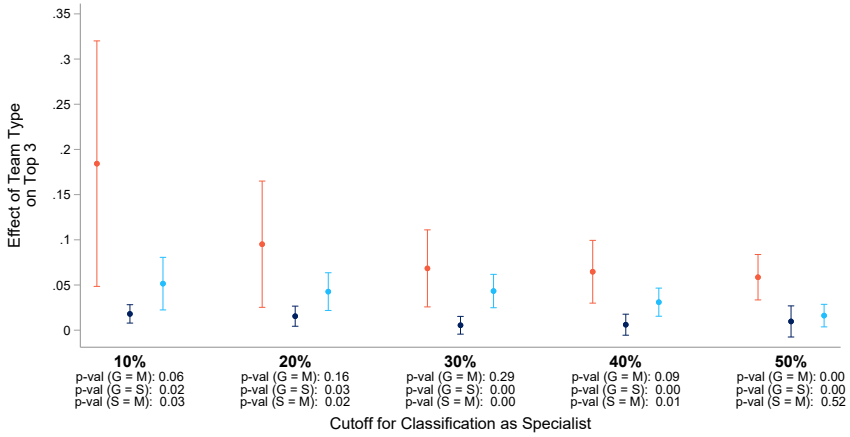
(a) Rank



(b) Medal Win



(c) Top 3



Notes: The figure plots the estimated coefficients τ_1, τ_2, and τ_3 from Equation (1.8) with all covariates and competition fixed effects when classifying generalists at different cutoffs of the specialization index. Panel (a) displays estimates for *Rank*, panel (b) for *Medal Win*, and panel (c) for *Top 3*.

which is equal to one if all code files written by user i include the same combination of packages. I classify users with below median PPD as generalists. Table 1.C.6 presents results using these new specialization classifications. When excluding all code files written after the introduction of ChatGPT, I do not find significant percentile rank differences between generalist and specialist teams anymore. However, generalist teams are still significantly more likely than specialist teams to win a medal or place among the top three. When using PPD as a measure of specialization, I find that generalist teams place about 8.5 percentile ranks higher than specialist teams and are 12 percentage points more likely to win a medal than specialist teams. I cannot detect significant differences for the probability of placing among the top three.

Overall, the results from Section 1.4.2, Section 1.4.3, and Section 1.4.4 support Prediction 1 as well as an interpretation of these differences as reflecting fundamental differences in the way generalist teams solve problems, rather than a different ability composition of generalist teams. The next sections turn towards mechanisms.

1.5 Mechanisms

My theoretical framework attributes quality differences between specialist and generalist teams to coordination costs. As I cannot measure coordination costs directly, I present several sets of diagnostic evidence: First, solution quality differences between generalist and specialist teams are driven by high complexity competitions. Second, the introduction of ChatGPT, a coordination-cost reducing technology, eliminates quality differences between generalist and specialist teams. Finally, I rule out four alternative explanations that could plausibly generate quality differences between team types: Generalists being more motivated and exerting more effort, generalists as better managers, generalists having higher social skills, and generalists being able to find better team members.

1.5.1 Investigating Complexity

A more complex problem requires more coordination, for example, because parts of the algorithm may interact in an unforeseen way, or because it may be less clear how to split the problem in sub-tasks. Prediction 2 echoes this intuition. As complexity increases coordination costs, quality differences between generalists and specialists should be larger for higher complexity problems. I use Prediction 2 as a first diagnostic to test whether solution quality differences between generalists and specialist arise because of higher coordination costs for generalist teams. To do so, I split the sample into high and low complexity competitions, and estimate Equation 1.8 separately for both types of competitions. I define a high-complexity competition as a competition with an overfitting risk in the top quartile of the overfitting distribution (c.f. Section 1.3.4.1). Figure 1.4 displays the

main coefficients and Table 1.D.7 full regression results. Generalist and specialist teams achieve similar ranks (Figure 1.4a), and are similarly likely to win a medal (Figure 1.4b) in low-complexity competitions. In high-complexity competitions however, generalist teams achieve higher quality than both mixed and specialist teams. For instance, in low-complexity competitions, generalist teams place only 2 percentile ranks higher than specialist teams on average, a difference that is not statistically significant, in high-complexity competitions however, the gap between generalist and specialist teams is equivalent to almost 10 percentile ranks and highly significant. In fact, all teams perform worse in high complexity competitions, but specialist teams are most affected by problem complexity. With the exception of *Top 3*, where generalist teams are best throughout, generalist teams only outperform specialist teams in high-complexity competitions.

Robustness To ensure my conclusions are not driven by how I classify complexity, I conduct several robustness checks. First, I change the complexity cut-off for a "high-complexity" competitions to the 50^{th} (Table 1.D.8) and to the 90^{th} percentile of competition complexity (Table 1.D.9). In both exercises, I only detect significant rank and medal differences between generalist and specialist teams in high-complexity competitions. Second, I use another measure of overfitting risk in a competition, competition *difficulty*, which is defined as 1− the share of teams whose best performing submission on the training dataset was also their best performing submission on the test dataset (Trotman, 2019). Again, I can only detect significant differences between generalist and specialist teams' solution quality in high-difficulty competitions. Finally, I assess problem complexity from two different angles, the length of instruction and whether the competition additionally specifies compute constraints. I extract competition descriptions from each competition's overview site on Kaggle. I consider a competition with above average instruction length a high complexity competition. Many competitions with the goal of detecting diseases from biomedical images fall under this category, like segmenting and grading biopsy images (c.f. Bulten et al., 2022), whereas competitions using tabular data with the goal to, for example, predict financial transactions, often have shorter instructions. Table 1.D.11 reports results. Here again, quality differences between generalist and specialist teams only emerge in high-complexity competitions. Last, a subset of Kaggle competitions specify additional compute constraints. To be eligible to win, competitors' algorithms have to run in a given time on given hardware. Competitors thus have to optimize along two dimensions, maximizing prediction accuracy while minimizing runtime, a more complex task. Table 1.D.12 displays results when splitting the sample into unconstrained and constrained competitions. As before, I only detect significant quality differences between generalist and specialist teams for more complex competitions, that is, in competitions *with* compute constraints.

1.5.2 ChatGPT as a Shock to Coordination Costs

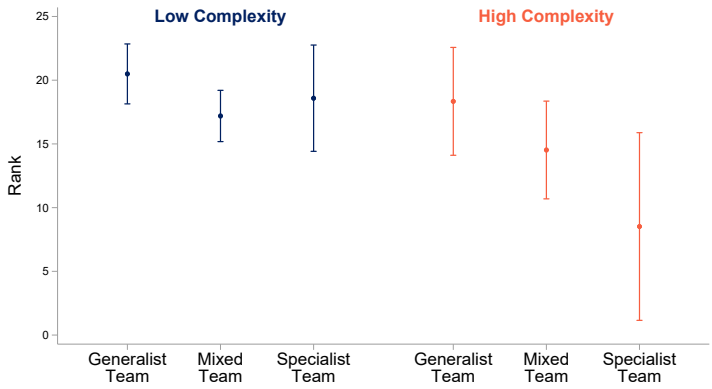
In the last section, I tested for coordination costs as a mechanism through an indirect shifter of coordination costs, problem complexity. Here, I will examine a direct shock to coordination costs: The introduction of ChatGPT. Technologies like ChatGPT could reduce coordination costs in teams in multiple ways, for example, by writing quick code to integrate components, by explaining code written by another team member, or by narrowing down the search space. At Google, large language model coding assistants were found to significantly speed up the code review process, one of the key elements of teamwork, by translating verbal code improvement suggestions made by other team members to code suggestions (Frömmgen et al., 2024). OpenAI introduced ChatGPT on November 30, 2022 as a way for users to interact with their large language model GPT-3. The introduction of ChatGPT was unexpected, both to the broader public (Figure 1.D.4a) and the Kaggle community, but attracted more attention on Kaggle than other large language models, like ChatGPT's predecessor GPT-2 or Google's BERT (Devlin et al., 2019) (Figure 1.D.4b and Figure 1.D.4c).

Similarly to the complexity analysis in the last section, I split the sample into competitions completed before November 30, 2022, and after. I assign competitions active during the introduction of ChatGPT to the post period, since all active competitions remained active for at least three weeks after the introduction of ChatGPT, enough time for competitors to use ChatGPT. I also truncate the pre-period to the exact same length as the post-period – my sample ends in February 2024 – to ensure that I am not attributing disappearing quality gaps a to lack of power.

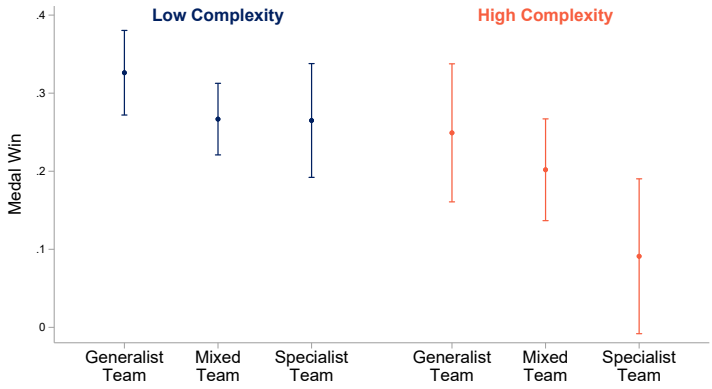
Figure 1.5 displays main coefficients and Table 1.D.13 full split-sample regression results. Before the introduction of ChatGPT, generalist teams place on average 11 percentile ranks higher than specialist teams, but after ChatGPT, specialist teams' ranks are indistinguishable from generalist teams', even suggesting they might outperform generalist teams. In the period before ChatGPT, specialist teams are not more likely to win a medal than solo competitors, and about 29 percentage points less likely than generalist teams to win a medal. This difference shrinks drastically in the post-ChatGPT period, and is not statistically significant anymore. I fail to detect significant differences in the probability of placing among the top three in both periods, likely due to substantially reduced sample sizes compared to Table 1.6.

Figure 1.4: Team Type, Complexity, and Solution Quality

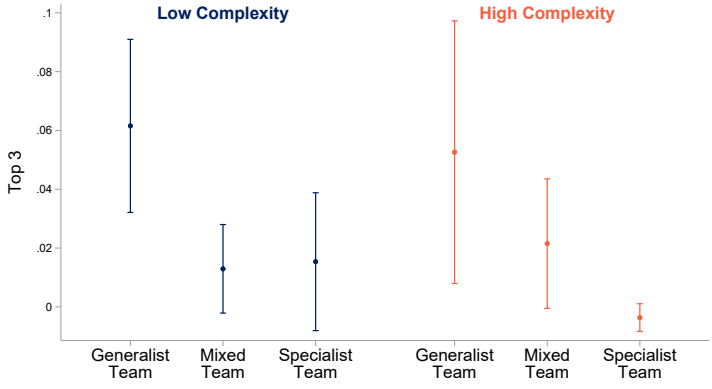
(a) Rank



(b) Medal Win



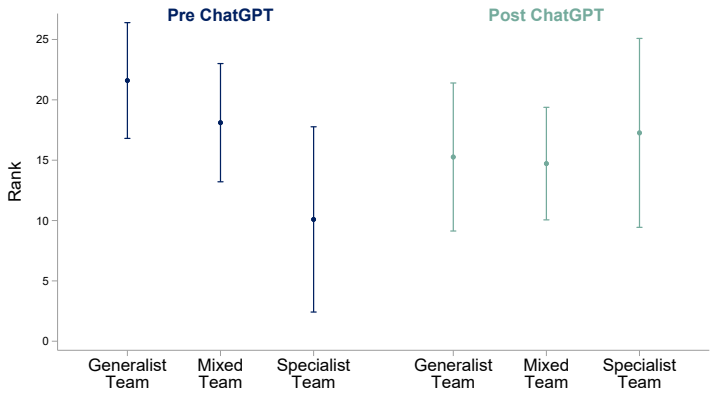
(c) Top 3



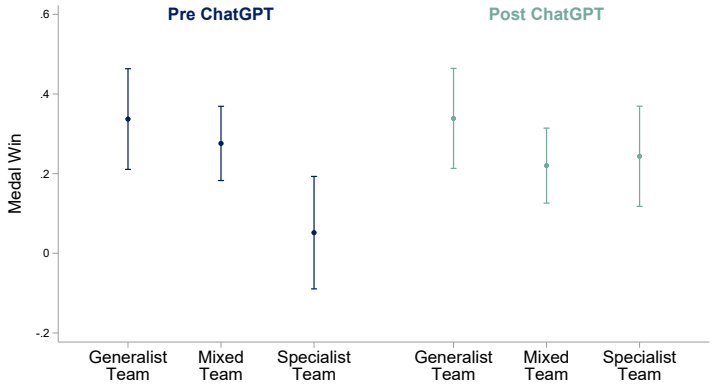
Notes: The figure plots the estimated coefficients τ_1, τ_2, and τ_3 from Equation (1.8). Regressions are estimated separately for the subset of low- and high-complexity competitions. A competition is classified as a high-complexity competition if its overfitting risk lies in the top quartile of the overfitting risk distribution. Panel (a) displays estimates for *Rank*, panel (b) for *Medal Win*, and panel (c) for *Top 3*. I report coefficients from regressions using all covariates and competition fixed effects as in column (5) in Tables 1.4-1.6. The corresponding regression results are reported in Table 1.D.7.

Figure 1.5: Team Type, ChatGPT, and Solution Quality

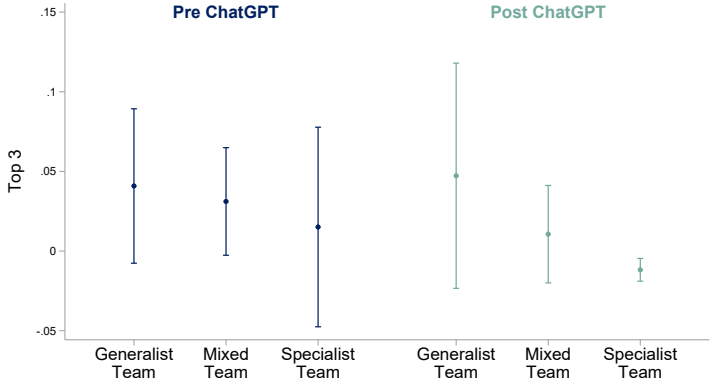
(a) Rank



(b) Medal Win



(c) Top 3



Notes: The figure plots the estimated coefficients τ_1, τ_2, and τ_3 from Equation (1.8). Regressions are estimated separately for competitions ending before and after the introduction of ChatGPT on November 30, 2022. Panel (a) displays estimates for *Rank*, panel (b) for *Medal Win*, and panel (c) for *Top 3*. I report coefficients from regressions using all covariates and competition fixed effects as in column (5) in Tables 1.4-1.6. The corresponding regression results are reported in Table 1.D.13.

1.5.3 Effort and Motivation

An alternative explanation for the documented quality differences between generalist and specialist teams is that, rather than these teams having different levels of coordination costs, generalist teams are more motivated and exert more effort. For instance, Weidmann and Deming (2021) find that individuals who increase team performance do so by raising the effort levels of their team peers. Generalists might take that role in their teams, for example, because they are able to keep the big picture in view and inspire their team with that vision. In Table 1.8, I investigate whether teams differ in terms of effort provision and motivation.

Column (1) presents results for Equation (1.8) with *Total Submissions* as a measure of effort. The number of submissions is highly correlated with performance in a competition. More submissions mean more feedback and more opportunities for fine-tuning the algorithm. Importantly, more submissions mean more work.[22] When controlling for ability proxies, team demographics, team size, code characteristics, and competition fixed effects, I do not detect significant differences in the number of total submissions across team types. Columns (2) and (3) report estimates for a finer dimension of motivation, the care taken to select the final submission. I define *Best Public Chosen* and *Best Private Chosen* as indicators whether a team chose the best performing submission on the public (known during the competition) or private (not known to teams during the competition) test set as a final submission. Note that Kaggle selects the best public submission as the final submission unless the team makes a different choice. Coefficients can thus be interpreted as making a conscious choice which requires more effort. I do not find any significant differences in the propensity to choose the best public or private submission between generalist teams, specialist teams and mixed teams.

Since rushing has been found to both be negatively correlated with effort and to reduce the quality of innovation (Weidmann and Deming, 2021; Hill and Stein, 2025), I examine two measures of rushing: *Time to Best*, the time it took a team to reach their highest ranked submission and *Time to Final* the time to reach the submission they chose as their final submission. On average, teams take longer than individual competitors to develop their best and final submissions. The interpretation of this difference is ambiguous. In part, this may explain team's better performance, since rushing a submission could be associated with lower quality (c.f. Hill and Stein, 2025). However, a longer time to submit might also reflect teams' higher coordination costs. While I do not detect any significant differences between team types in the time that it takes a team to develop their best performing submission, generalist teams take slightly longer than mixed teams to develop their final

[22] Kaggle enforces a daily submission limit, implicitly censoring effort measures. Kaggle's rationale behind this is to ensure a level playing field – larger teams might have more capacities to submit more and receive more feedback than smaller teams. However, the overwhelming majority of teams submit less than the limit specified, see Figure 1.D.5.

Table 1.8: Team Type and Submission Behavior

Dependent Variable:	Total Submissions	Chose Best Public	Chose Best Private	Time to Best	Time to Final
	(1)	(2)	(3)	(4)	(5)
Generalist Team	43.85***	-0.12***	-0.17***	0.13***	0.15***
	(4.32)	(0.02)	(0.02)	(0.01)	(0.01)
Mixed Team	40.18***	-0.15***	-0.20***	0.11***	0.12***
	(3.03)	(0.02)	(0.02)	(0.01)	(0.01)
Specialist Team	47.80***	-0.13***	-0.19***	0.12***	0.13***
	(5.34)	(0.03)	(0.03)	(0.02)	(0.02)
Competition Experience	-0.08***	-0.00	0.00	0.00***	0.00***
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.29***	-0.00***	-0.00***	0.00***	0.00***
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
3-Person Team	41.16***	0.01	-0.01	0.04***	0.04***
	(6.22)	(0.03)	(0.03)	(0.01)	(0.01)
4-Person Team	80.34***	0.01	-0.12**	0.08***	0.10***
	(13.25)	(0.06)	(0.05)	(0.02)	(0.02)
5-Person Team	116.20***	-0.10*	-0.16***	0.10***	0.12***
	(13.83)	(0.06)	(0.06)	(0.02)	(0.02)
R^2	0.18	0.11	0.16	0.22	0.22
Observations	60090	60090	60090	60090	60090
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.48	0.38	0.20	0.13	0.06
p-value (Generalist Team = Specialist Team)	0.56	0.93	0.55	0.58	0.37
p-value (Specialist Team = Mixed Team)	0.20	0.54	0.69	0.63	0.75

Notes: The table reports the estimates of equation (1.8). The dependent variable is a team's total number of submissions (1), an indicator whether a team chose the submission that performed best in the public test sample as their final competition submission (2), an indicator whether a team chose the submission that performed best in the private test sample as their final competition submission (3), the time the team took to develop their best submission (4), the time the team took to develop their final competition submission (5). I measure time in the fraction of days relative to total competition duration, since the value of a day is different if the competition lasts two weeks or two months. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Competition experience measures the average of the number of competitions each team member participated in before competition c. Ability controls include Previous Rank (the average percentile rank achieved by each team member when competing alone in previous competitions) and Experience (the average number of competitions that each team participated in on their own in previous competitions). Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

submission. For example, in a 100-day competition, generalist teams develop the submission that they select as their final competition submission three days later than mixed teams. Overall, I find little support for differential effort as a mechanism behind generalists-specialist quality differences.

1.5.4 Management Skills

Recent empirical work has highlighted the importance of managers for allocating workers to specialized tasks and thereby increasing productivity (Minni, 2024; Weidmann et al., 2024). Consistent with theoretical arguments made by Crémer et al. (2007), the role of generalists in teamwork could also be that of a manager. In Crémer et al. (2007), different units of an organization engage

in solving parts of a problem. However, they do not "speak" the same language, i.e., they are specialized on different parts of the problem. A manager, who "speaks" both languages, assigns the parts of the problem to the person who can solve them. An interpretation of their theoretical model is that individuals with broader knowledge make for better managers.

Although there are no explicit managers in Kaggle teams, each team is required to select a team leader. The team leader has de facto authority: They are the primary point of contact between the platform and the team, have the sole authority over accepting new team members and merging teams, and make the final submission selections.[23] These powers and responsibilities align closely with a management role. To investigate whether generalists are better managers, I estimate the following regression in the sample of mixed teams[24]:

$$\mathrm{Y}_{tc} = \alpha + \beta_1 \text{Generalist Team Leader}_{tc} + X'_{tc}\beta + \epsilon_{tc} \tag{1.11}$$

where Generalist Team Leader$_{tc}$ is an indicator that is equal to one if team t's team leader is a generalist, and Y_{tc} and X_{tc} are defined as in Equation (1.8).

Table 1.9 presents results. I do not find any evidence that teams with a generalist team leader perform any different than teams with a specialist leader. In columns 4-6, I additionally split ability controls into the ability of the team leader and that of the other members, and control for both separately. The effects of team lead and team member ability are indistinguishable (F-Tests: 0.31 (Column 4), 0.34 (Column 5), 0.32 (Column 6)).

1.5.5 Social Skills

Social skills are crucial for teamwork (Deming, 2017; Weidmann and Deming, 2021; Adhvaryu et al., 2023). For example, Deming (2017) explicitly models social skills as a factor that reduces coordination frictions between specialized workers. Social skills could be correlated with the degree of specialization, for example, if a broader set of interests is correlated with an open mind towards people. Any quality difference between generalist and specialist teams would then be a result of social skills rather than coordination costs differing by the degree of specialization. Do generalist teams have higher levels of social skills? Since I do not have access to explicit measures of social skills, I analyze users' public communication behavior on the platform to capture social skills. Kaggle users can engage in forum discussions. Aside from some platform-wide forums for questions or product feedback, each competition has an attached discussion forum in which users can and very frequently do interact. From these messages, I construct several metrics of social skills, and investigate whether team types differ in their levels of social skills. Note that I do not have access to private, within team communication data. A necessary assumption is thus that interactions with

[23] See Kaggle.com (2024a) and Reade (2024).
[24] As only generalists can be team leaders in generalist teams and analogously for specialists.

Table 1.9: Generalist Team Leaders in Mixed Teams

Dependent Variable:	*Rank*	*Medal Win*	*Top 3*	*Rank*	*Medal Win*	*Top 3*
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team Leader	-0.26	-0.03	-0.01	0.14	-0.06	-0.01
	(1.69)	(0.04)	(0.01)	(1.71)	(0.05)	(0.02)
Competition Experience	0.15**	0.00**	-0.00			
	(0.07)	(0.00)	(0.00)			
Previous Rank	0.29***		0.00			
	(0.08)		(0.00)			
Previous Medal Wins		0.51***				
		(0.17)				
Team Lead's Experience				0.02	0.00	-0.00
				(0.04)	(0.00)	(0.00)
Team Members' Experience				0.08*	0.00	-0.00
				(0.04)	(0.00)	(0.00)
Team Lead's Rank				0.18***		0.00**
				(0.06)		(0.00)
Team Members' Rank				0.09		0.00
				(0.07)		(0.00)
Team Lead's Medal Wins					0.49***	
					(0.12)	
Team Members' Medal Wins					0.31**	
					(0.12)	
3-Person Team	5.06**	0.15***	0.03	7.35***	0.20***	0.05*
	(2.38)	(0.05)	(0.02)	(2.10)	(0.06)	(0.03)
4-Person Team	12.48***	0.26***	0.04	10.83***	0.25***	0.04
	(2.32)	(0.06)	(0.03)	(2.33)	(0.06)	(0.04)
5-Person Team	13.19***	0.27***	-0.02	12.02***	0.32***	-0.03
	(2.59)	(0.07)	(0.03)	(2.56)	(0.07)	(0.04)
R^2	0.56	0.52	0.40	0.62	0.57	0.45
Observations	830	830	830	655	655	655
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports the estimates of equation (1.11). The dependent variable is a team's percentile rank, (1) and (4), an indicator whether a team won a medal, (2) and (5), and an indicator whether a team placed among the Top 3, (3) and (6). The main explanatory variable is an indicator whether the team leader is a generalist. Competition experience measures the average of the number of competitions each team member participated in before competition c. Ability controls include Previous Rank (the average percentile rank achieved by each team member when competing alone in previous competitions) and Experience (the average number of competitions that each team participated in on their own in previous competitions). Team Lead and Team Members' variables are defined in the same way, but included disaggregated by team leader and non-team leader team members. Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

a larger group of peers generalize to behavior within a smaller group of peers. This is plausible in this context, since public communication is not anonymous. I estimate the following regression on the sample of stable teams:

$$\mathrm{Y}_t = \alpha + \tau_1 \text{Generalist Team}_t + \tau_2 \text{Mixed Team}_t + \tau_3 \text{Specialist Team}_t + X_t'\beta + \epsilon_i \tag{1.12}$$

where Y_t is a measure of the team's average social skills, Generalist Team_t, Mixed Team_t, and Specialist Team_t are defined as before, and X_t is a vector of team-level demographics and team size indicators. I first evaluate whether the total volume of messages differs between team types. Engaging more with public message boards is an indicator for a desire and willingness to communicate with others, a prerequisite of social skills. I do not find any significant differences between team

Table 1.10: Team Type and Social Skills

Dependent Variable:	Number Messages	Share with Thanks	Thanks per Message	Share with User Mentions	User mentions per Message	Smileys per Message
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	139.58***	-0.07***	-0.07***	0.03***	0.04***	0.02***
	(16.12)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Mixed Team	119.59***	-0.05***	-0.05***	0.04***	0.05***	0.01***
	(11.85)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Specialist Team	140.73***	-0.03**	-0.03**	0.05***	0.05***	0.03***
	(30.84)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
3-Person Team	105.48***	-0.02*	-0.02	0.02**	0.04***	0.00
	(25.07)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4-Person Team	283.99***	-0.04***	-0.03**	0.04**	0.06***	-0.00
	(81.83)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
5-Person Team	203.89***	-0.03**	-0.03**	0.04***	0.06***	-0.00
	(35.67)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
R^2	0.05	0.03	0.03	0.05	0.03	0.03
Observations	13030	13030	13030	13030	13030	13030
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.30	0.03	0.02	0.22	0.30	0.62
p-value (Generalist Team = Specialist Team)	0.97	0.01	0.01	0.13	0.41	0.09
p-value (Specialist Team = Mixed Team)	0.51	0.21	0.20	0.46	0.93	0.04

Notes: The table reports the estimates of equation (1.12). The dependent variables are the average number of forum messages posted by a team's members (1), the average share of their messages that contain expressions of gratitude (2), as well as the share of these per message (3), the average share of messages mentioning other users (4), as well as the share of user mentions per message (5), and the average share of smileys per message (6). The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

types (Table 1.10, Column 1). I next turn to the content of the messages: Experimental research from psychology indicates that explicit expressions of gratitude help initiate social relationships (Williams and Bartlett, 2015). I thus investigate what portion of a user's forum posts include "thank you's" and how many "thank you's" an average message contains (Table 1.10, Columns 2 and 3). Specialist Teams are comprised of users who thank others more often. I also examine how often other users are mentioned in messages and how often (Table 1.10, Columns 4 and 5)[25] as a more direct measure of social ties. I do not find any significant differences. Finally, I construct a simple measure of friendliness: whether a message includes graphical representations of emotion, i.e., smileys.[26] Table 1.10, Column 6 presents results. Among all teams, specialist teams include most smileys in their messages. Taken together, I do not find striking differences in various measures of social skills between team types. Since specialist teams appear at least as, if not more, sociable as generalist teams, social skill differences are unlikely to explain the latter's superior performance.

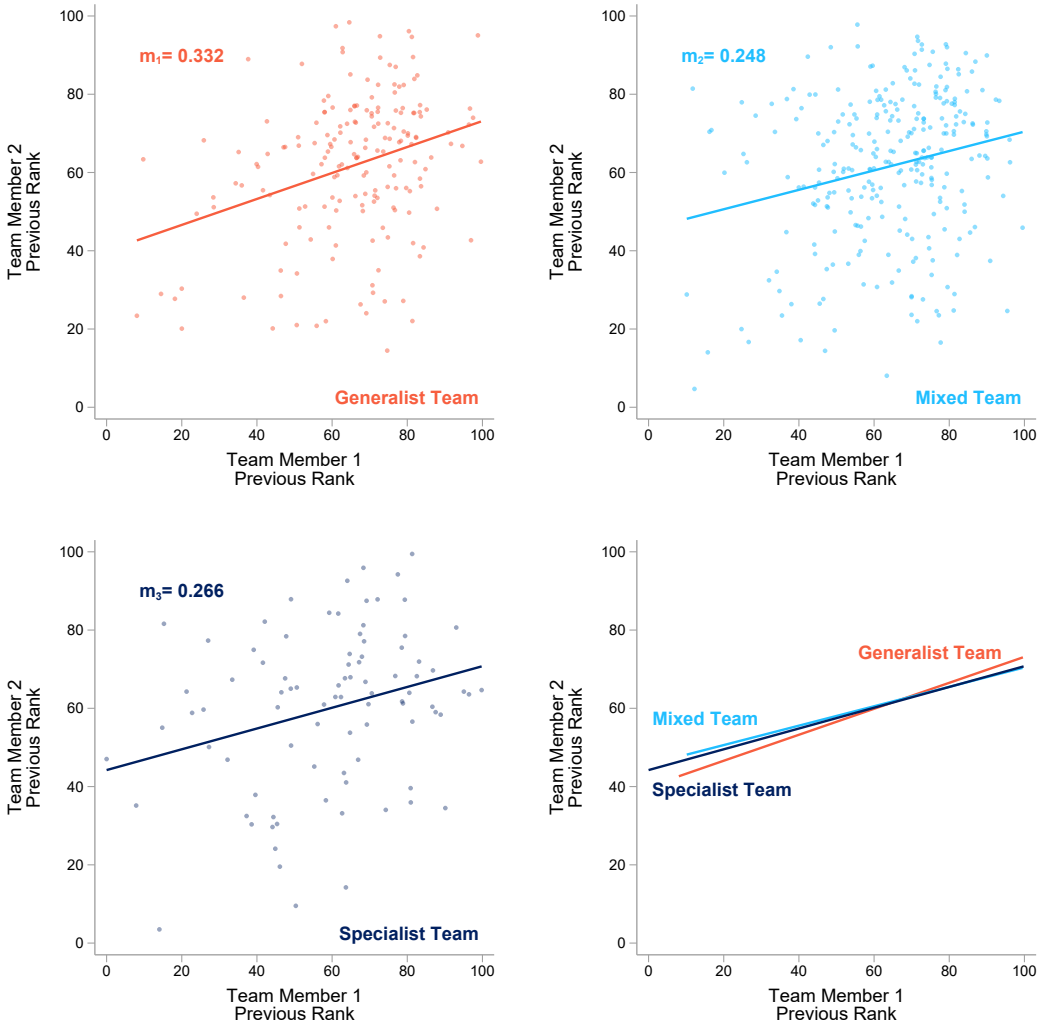
1.5.6 Team Member Matching

For the main part of this paper, my focus has been on differing *costs* of teamwork between generalists and specialists. However, the productivity *gains* from collaboration might also differ. Both the literature on peer effects (e.g., Hamilton et al., 2003; Mas and Moretti, 2009) and macroeconomic

[25] References to other users are clearly identifiable through HTML tags in the unprocessed message data which I have access to.

[26] Specifically, I check for the presence of the following character sequences: *:) =) ;) ;-) :-) (: (= (; (-; (-:* and the unicode strings for several graphical emoticons.

Figure 1.6: Assortative Matching in Different Team Types



Notes: The figure shows the correlations between team member's previous percentile ranks in two-person teams. Team member 1 is the team leader. Panel (a) displays generalist teams, panel (b) mixed teams, and panel (c) specialist teams. m_1, m_2, and m_3 are coefficients from a regression of team member 2 previous rank on team member 1 previous rank. Panel (d) compares the correlations across team types.

models of labor market inequality (Cornelissen et al., 2017; Jarosch et al., 2021; Herkenhoff et al., 2024; Freund, 2024) stress coworker complementarities in teams, finding that workers' productivity disproportionately increases when matched with higher-ability coworkers. Although it is unclear whether and to which extend these complementarities exist in science – Azoulay et al. (2010) find positive, Waldinger (2011) no, and Ahmadpoor and Jones (2019) negative productivity spillovers – we might still be worried that specialists and generalists sort into teams with different spreads in ability. For instance, generalist teams might systematically be more heterogeneous in ability,

because generalists' broader knowledge might allow them to assess others' abilities better and find higher ability team members given their own ability. If one high ability team member lifts the others up disproportionately, a generalist team with a larger heterogeneity in ability would produce a higher quality solution than a specialist team with the same average ability.

To explore whether different ability matching patterns could explain generalist-specialist quality differences, I examine the correlation between team members' previous performance in Figure 1.6. I focus on a two-person team's first collaboration, and designate the team leader as team member 1. The first three panels in Figure 1.6 show scatter plots of teams' ability combinations and the association between team members' ability as a linear fit, separately for each team type. I find similar correlations between the two team members' previous performance for all team types: When one team member has previously placed 10 percentile ranks higher, the other team member's previous performance increases between 2.5 to 3.3 percentile ranks. These associations are statistically indistinguishable. The p-value of the Chi2-Test comparing the differences in slopes between the sample of generalist teams and mixed teams is 0.42, 0.57 comparing generalist and specialist teams, and 0.88 for specialist and mixed teams. Given their own ability, generalists do not seem to be able to find better team members than specialists, making it unlikely that differing heterogeneity of team ability drives quality differences of specialists and generalists.

1.6 Conclusion

Making teamwork succeed is crucial for maintaining high rates of innovation in a time in which ideas are getting harder to find (Bloom et al., 2020) and innovation is increasingly produced in teams (Wuchty et al., 2007). In this paper I show that there is a tension between specialization and teamwork, as generalist teams produce higher quality innovation than specialist teams. I argue that this is a result of coordination costs, which are higher in specialist teams. Using data from online machine learning competitions, I employ several empirical strategies to disentangle the effect of team type from the contributions of individual team members. I consistently find that generalist teams produce higher quality than specialist teams. I provide several pieces of evidence that point towards coordination costs as a channel. When working on a low-complexity problem, specialist teams perform at par with generalist teams. Innovation quality differences only arise in high-complexity problems, for which coordination costs are particularly high. A shift in coordination costs due to the introduction of ChatGPT also removes quality differences between generalists and specialists. I find no evidence for alternative mechanisms, such as motivation, management skills, social skills, or collaborator matching, which might favor generalist teams. Overall, my findings highlight that we need to take skill *breadth* into account if we want to build innovative

teams. Universities and research agencies invest considerable funds into programs designed to foster interdisciplinary research. While previous research suggests that teams from more diverse disciplinary backgrounds produce higher-impact innovation (Uzzi et al., 2013; Allocca, 2024), I find that coordination costs hamper the productivity of such teams and might lead these endeavors to fail.

Appendix to Chapter 1

- Appendix 1.A provides further details and derivations for the theoretical framework in Section 1.2.
- Appendix 1.B provides further details on the construction of the data.
- Appendix 1.C reports robustness checks and additional findings related to Section 1.4.
- Appendix 1.D reports robustness checks and additional findings related to Section 1.5.

1.A Model: Derivations

To derive the predictions in Section 1.2.2, I solve the model outlined in Section 1.2 via backward induction, first deriving optimal effort, then optimal task division, and finally the conditions for collaborating.

Effort Choices and the Division of Tasks within Teams When working alone, a scientist has to execute all tasks by themselves, meaning that Q_p^a and ${e_i^a}^*$ are zero when the problem falls outside of the range of tasks a scientist can solve, i.e., $\Theta_p > b_i$.
Else, optimal efforts are given by:

$$\frac{\partial q(\Theta_p, \sigma_i, e_i^{a*})}{\partial e_i^{a*}} = \Theta_p \frac{\partial c({e_i^a}^*)}{\partial e_i^{a*}} \tag{1.A.1}$$

As $\partial q(w_i, \sigma_i, e_i)/\partial e_i \partial \sigma_i > 0$ and $\partial c(e_i)/\partial \sigma_i = 0$:
Corollary 1: Of two scientists working alone on the same problem Θ_p, the more specialized scientist (higher σ_i) will exert higher effort. Conversely, more specialized scientists will work on fewer problems alone.
Looking at team production, let us first consider the case where both scientists have $b_i \geq \Theta_p$. Optimal efforts are given by:

$$\frac{1}{2}\frac{\partial q(w_i, \sigma_i, e_i^{t*},)}{\partial e_i^{t*}} = w_i \frac{\partial c(e_i^{t*})}{\partial e_i^{t*}} \tag{1.A.2}$$

Since Equation (1.A.2) has to hold for both team members, the optimal division of labor within a team is given by:

$$\frac{\partial q(w_i, \sigma_i, e_i^{t*},)}{\partial e_i^{t*}} \Big/ \frac{\partial q(w_j, \sigma_j, e_j^{t*},)}{\partial e_j^{t*}} = \frac{w_i}{w_j} \frac{\partial c(e_i^{t*})}{\partial e_i^{t*}} \Big/ \frac{\partial c(e_j^{t*})}{\partial e_j^{t*}} \tag{1.A.3}$$

meaning, two equally specialized scientists will execute exactly half of the necessary tasks.
In an homogeneous team ($\sigma_i = \sigma_j$), optimal efforts are then given by:

$$\frac{\partial q(\Theta_p/2, \sigma_i, e_i^{t*})}{\partial e_i^{t*}} = \Theta_p \frac{\partial c(e_i^{t*})}{\partial e_i^{t*}} \tag{1.A.4}$$

Due to $\partial q(w_i, \sigma_i, e_i)/\partial e_i \partial w_i < 0$, for homogeneous teams, individual efforts per task in a team are larger than when working alone. In an mixed team ($\sigma_i \neq \sigma_j$), the more specialized scientist executes a larger share of tasks to equalize production across the team. Denote as α the share of tasks taken over by the more specialized scientist. Since $\alpha > 1/2$, the change in effort for the more specialized scientist is ambiguous. Efforts only increase if $\partial q(w_i, \sigma_i, e_i)/\partial e_i \partial w_i < 0$ is sufficiently

small. However, efforts of the less specialized scientist always increase.

Corollary 2: In a team of two equally specialized scientists, both scientists will exert more effort per tasks than when working alone. In a team of two scientists with different levels of specialization, the less specialized scientist will exert more effort, whereas the change in effort for the more specialized scientist is ambiguous. Homogeneous teams exert more effort than individuals for the same problem, mixed teams the same or more. Now, what happens when two scientists collaborate, and one of them has $b_j < \Theta_p$? Clearly, the less specialized scientist now has to execute tasks $\Theta_p - b_j$, and potentially an additional share $(1-\varphi)$ of the tasks $\theta < b_j$. Equation (1.A.2) becomes:

$$\frac{\partial q(\Theta_p - \varphi b_j, \sigma_i, e_i^{t*})}{\partial e_i^{t*}} \Big/ \frac{\partial \varphi b_j, \sigma_j, q(e_j^{t*},)}{\partial e_j^{t*}} = \frac{\Theta_p - \varphi b_j}{\varphi b_j} \frac{\partial c(e_i^{t*})}{\partial e_i^{t*}} \Big/ \frac{\partial c(e_j^{t*})}{\partial e_j^{t*}} \tag{1.A.5}$$

Since we know that in an efficient team, the more specialized scientist executes a larger share of tasks, a team with $\Theta_p/2 \geq b_j$ will never be efficient. In this case, scientist j executes b_j tasks, and scientist i the rest.

Collaboration Choices Recall that, for a given problem Θ_p, a scientist σ_i collaborates if:

$$\frac{1}{2}\underbrace{\{Q_p^t - C(\gamma_p, \sigma_i, \sigma_j)\}}_{\text{team solution quality}} - \underbrace{w_i c(e_i^{t*})}_{\text{total cost of effort}} \geq \underbrace{Q_p^a}_{\text{individual solution quality}} - \underbrace{\Theta_p c(e_i^{a*})}_{\text{total cost of effort}} \tag{1.A.6}$$

where e_i^{a*} and e_i^{t*} are optimal efforts when working alone or working as a team. For a homogeneous team, $(\sigma_i = \sigma_j)$ Equation (1.A.6) becomes:

$$\underbrace{\frac{1}{2}Q_p^t - \frac{\Theta_p}{2}c(e_i^{t*}) - Q_p^a + \Theta_p c(e_i^{a*})}_{\text{return to collaboration}} \geq \underbrace{\frac{1}{2}C(\gamma_p, \sigma_i, \sigma_i)}_{\text{collaboration costs}} \tag{1.A.7}$$

That is, the return to collaboration has to be at least as large as the collaboration costs borne by one team member.

A mixed team collaborates if:

$$\underbrace{\frac{1}{2}Q_p^t - \alpha\Theta_p c(e_i^{t*}) - Q_p^a + \Theta_p c(e_i^{a*})}_{\text{return to collaboration for more specialized scientist}} \geq \underbrace{\frac{1}{2}C(\gamma_p, \sigma_i, \sigma_j)}_{\text{collaboration costs}} \tag{1.A.8}$$

for the more specialized scientist. If Equation (1.A.8) holds for the more specialized scientist, it also holds for the less specialized scientist, because working alone is even less profitable for the less specialized scientist. Let's look at the case where one scientist has $b_i \geq \Theta_p$. Clearly, they will never collaborate with a scientist who has the same level of specialization, since they will not be able to execute all tasks necessary to solve the problem. In this scenario, when will it be profitable for a

less specialized scientist to collaborate with that more specialized scientist?

$$\underbrace{\frac{1}{2}Q_p^t - (\Theta_p - \varphi b_j)c(e_i^{t*}) - Q_p^a + \Theta_p c(e_i^{a*})}_{\text{return to collaborating for less specialized scientist}} \geq \underbrace{\frac{1}{2}C(\gamma_p, \sigma_i, \sigma_j)}_{\text{collaboration costs}} \tag{1.A.9}$$

Here, the relevant constraint is on the less specialized scientist: The more specialized scientist's return from working alone is zero.

Are all teams equally likely to occur? Due to lower coordination costs, collaboration between two scientists with low levels of specialization is easier to sustain than between to more specialized scientists if $\partial q(w_i, \sigma_i, e_i)/\partial w_i \partial \sigma_i = 0$, i.e., the gains from the division of labor are the same across all levels of specialization. If $\partial q(w_i, \sigma_i, e_i)/\partial w_i \partial \sigma_i > 0$, which team is easier to sustain for a given problem depends on the relative strength of the increase in coordination costs versus the increase in quality. Assuming that scientists and problems are distributed i.i.d. over the unit interval, a mixed team will be most frequent, since there is a mass of problems that some specialists can only solve in collaboration, and these specialists will only collaborate with less specialized scientists.

Corollary 3: Not all types of collaborations are equally frequent: Specialist Teams are least likely, followed by generalist teams, and mixed teams. Specialists are less likely than generalists to work on a problem on their own.

Solution Quality Consider a case in which constraints 1.A.7-1.A.9 are fulfilled for a range of σ_i, σ_j pairs. Which team will produce the highest quality solution? Recall that solution quality for teams is:

$$Q = \min_{0 \leq \Theta_p} q(\theta) - C(\gamma_p, \sigma_i, \sigma_j) \tag{1.A.10}$$

Since efforts are chosen to produce the exact same quality across all sub-tasks, we can re-write team solution quality as:

$$Q^t = q(w_i^*, \sigma_i, e_i^{t*}) - C(\gamma_p, \sigma_i, \sigma_j) \tag{1.A.11}$$

Applying the envelope theorem to this expression:

$$\frac{\partial Q^t}{\partial \sigma_i} = \frac{\partial q}{\partial \sigma_i} - \frac{\partial C}{\partial \sigma_i} \tag{1.A.12}$$

yields Prediction 1: A generalist team is best when coordination costs are more elastic to specialization than task quality. Conversely, a reduction in the slope of coordination costs will shrink

performance gaps. For Prediction 2, we can differentiate Equation (1.A.12) with respect to γ_p:

$$\frac{\partial^2 Q}{\partial \sigma_i \, \partial \gamma_p} = -\frac{\partial^2 C}{\partial \sigma_i \, \partial \gamma_p} \tag{1.A.13}$$

Since coordination costs increase in complexity, i.e., $\partial C / \partial \gamma_p \partial \sigma > 0$ the difference in solution quality between more and less specialized teams also increases in complexity.

1.B Further Details on Data Construction

1.B.1 Competitions

Table 1.B.1: Kaggle Competiton Examples

Name	Hosted By	Description	Start Date	Duration (Days)	Total Prizes (1000 USD)	Competitors
Panel A: Featured Competitions						
Passenger Screening Algorithm Challenge	Department of Homeland Security	detection and classification of threats from TSA body scans	Jun 22, 2017	176	1500	217
Vesuvius Challenge - Ink Detection	Vesuvius Challenge	detection of ink on x-rays of papyri buried in the eruption of vesuvius	Mar 15, 2023	91	1000	1514
LLM - Detect AI Generated Text	The Learning Agency Lab	classification of essays into student or large language model written	Oct 31, 2023	83	110	5264
RSNA Screening Mammography Breast Cancer Detection	Radiological Society of North America	detection of breast cancers in screening mammogram images	Jan 28, 2022	91	50	2146
Panel B: Research Competitions						
Google - American Sign Language Fingerspelling Recognition	Google	detection and transcription of signed ASL characters from video	May 10, 2023	106	200	1530
BirdCLEF 2023	Cornell Lab of Ornithology	detection and classification of bird calls in audio data	Mar 7, 2023	78	50	1397
IceCube - Neutrinos in Deep Ice	IceCube Neutrino Observatory	reconstruction of the direction of neutrinos from sensor data	Jan 19, 2023	90	50	901
OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction	Stanford University	prediction of degradation points in mRNA molecules	Sep 11, 2020	25	25	1839

1.B.2 User Demographics

Coding User Gender I code user gender into two categories: female and male.[27] To do so, I proceed in four steps.

1. If a user reports pronouns, I use pronouns to assign gender (she=female, he=male). Only 1.11% of users report pronouns.

I next use user *display names*, i.e., the name that is shown on a user's profile, to assign gender. After cleaning names from numbers, transcribing non-roman alphabets (mostly Chinese, Japanese, Korean, and Russian) as well as removing symbols and emojis, and splitting names into first and last names, I assign gender by:

2. I retrieve information on the most likely gender for each name and country from gender-api.com. I only code a name as male or female if the probability of a name being associated with one gender is higher than 85%. If no country is reported, I assign a gender only if the gender most associated with a name does not change across countries. I can assign gender for 45.73% of users in this step.

3. If a name is not present in gender-api.com and reported at least twice, my research assistant and I try to determine the name's gender by a google search.

4. For the remaining names, which are often nicknames like "computervisionjedi" or "aimuaddib"[28], my research assistant hand-checked the user's profile for additional information such as profile pictures to code user gender.

Overall, I can assign a gender for 79.51% of users in my sample.

Coding User Occupation I use information provided by Kaggle users on their profile page to code occupation. I extract occupations from two fields: Occupation and Bio. I consider both these fields for three reasons: One, occupation is not always filled, two, contextual clues are necessary to disambiguate occupations like "architect", which could both refer to the traditional use of the term – someone who designs and builds houses – as well as the role of a data science professional, and three, some users report occupation aspirations as their occupation, but details in their bio reveal that they are not currently active in that occupation. For example, some users report "machine learning engineer" as their occupation, but write in their bio that they are a college junior. Some users also do not report occupation but their position in a firm's hierarchy, e.g., CEO at startup,

[27] Since users are able to report their preferred pronouns, I originally included a non-binary category. However, no user in my sample reports they/them as preferred pronouns.

[28] For data protection, these are not real Kaggle user names by Kaggle users, but composites from popular references.

but provide more detail in their bios. Table 1.B.2 displays chosen occupational categories, as well as the size of a given category and examples for reported occupations within that category. Category "Other" includes both users who do not report occupation information or who report an occupation that is not classifiable within these categories, for example, "teacher", "farmer", or "photographer".

Table 1.B.2: Occupation Categories and Examples

Occupation Category	Observations	%	Modal Occupation	Examples
Academic	125	0.92	Academic at -	Postdoc at University College London Professor in Urban Water Infrastructure at TU Delft
Analyst	243	1.78	Analyst at Deloitte	Analyst at OECD Business Research Analyst-I at Amazon
Artificial Intelligence Professional	160	1.17	AI Engineer	AI Engineer at IBM AI Vision Engineer at Samsung Electronics
Business Professional	202	1.48	Manager at Accenture	Manager at American Express Marketing Manager at Honda Motor Co. Project Manager at JP Morgan
Computer Science Professional	153	1.12	Research Engineer	Computer Systems Engineer at Lawrence Berkeley National Labs Engineering Manager at Adobe
Consultant	183	1.34	Consultant at EY	Analytics Consultant at Deloitte Partner at Mckinsey & Company
Data Analyst	270	1.98	Data Analyst at Home	Aircraft Data Analyst at Airbus Data and Analytics at H&M
Data Engineer	152	1.11	Data Engineer at Accenture	Dat Engineer at Paypal Inc Data Engineer at Google
Data Scientist	2233	16.36	Data Scientist at Freelance	Data Scientist at Uber Datascience at Telecom DS at H2O.Ai
Developer	100	0.73	Developer at Kaggle	Deleloper at Intel Python Developer at Infosys
Engineer	351	2.57	Engineer at JTC	Civil Engineer Engineer at BMW Group
Machine Learning Engineer	713	5.22	ML Engineer	Applied ML at Nvidia Cloud Solution Architect Data and AI at Microsoft Machine Learning Engineer at Apple
Master Student	206	1.51	Master	M.S. Canditate at ETH ZüRich MSBA at Ut Austin
Phd Student	338	2.48	Graduate Student at Carnegie Mellon University	Ph.D. Candidate in Economics at University Of Chicago CS Phd Student at UC Irvine Phd Student at Columbia University
Researcher	380	2.78	Research Scientist	Applied Scientist at Microsoft Computational Scientist at IBM Quantum Data Scientist & Researcher at Nvidia
Software Developer	696	5.10	Software Engineer	Algorithm Engineer at Alibaba SDE at Amazon Software Engineer at Bank Of America
Undergraduate Student	2369	17.36	Student	B.Sc Data Science Student Computer Science Undergraduate
Other	4772	34.97		

Notes: The table displays all occupation categories used as covariates in Equation (1.8). Spelling of modal occupation and examples correspond to original data.

1.B.3 Example Notebook

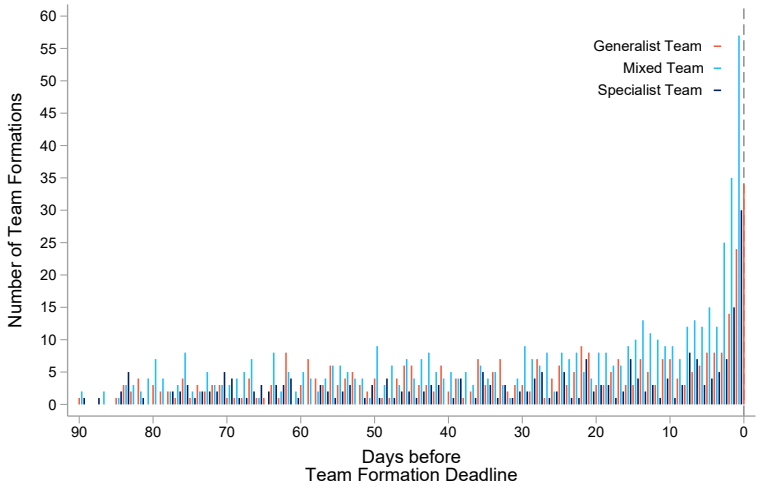
Figure 1.B.1: Kaggle Code Notebook Example

Note: Example adapted from Howard (2024).

1.C Additional Results for Section 1.4

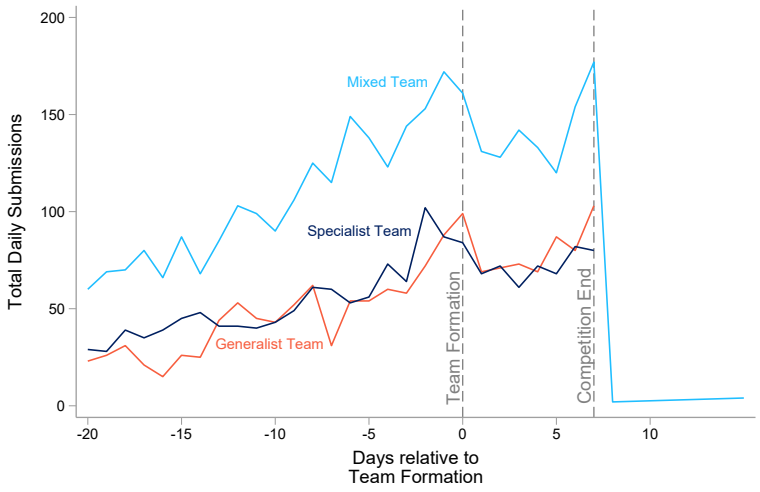
1.C.1 Additional Results for Section 1.4.4

Figure 1.C.2: Team Formations per Day



Notes: The figure plots the average number of team formations for generalist, mixed, and specialist teams in the three months leading up to the team formation deadline. Team formations are zero after the deadline by definition.

Figure 1.C.3: Total Daily Submissions



Notes: The figure plots aggregate daily submission counts for generalist, mixed, and specialist teams in the three weeks leading up to and the week after the team formation deadline. The dashed gray line indicates the modal competition end date at seven days after the team formation deadline.

Table 1.C.3: Event Study Coefficients and F-Tests for Figure 1.2

Days relative to Team Formation	before				after			
	more than 8	7-8	5-6	3-4	0-1	2-3	4-5	more than 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Generalist Team	1.02	3.17	-0.16	3.48	6.98**	8.52*	12.28***	13.08***
	(4.58)	(5.07)	(3.75)	(4.01)	(3.08)	(4.92)	(4.22)	(4.59)
Specialist Team	-2.61	-3.98	-4.68	-5.41	-1.09	-0.59	0.85	2.79
	(5.45)	(6.12)	(5.66)	(4.95)	(3.26)	(3.20)	(3.72)	(3.00)
Mixed Team	-1.93	0.20	0.10	2.43	4.69	5.58*	9.53***	7.36**
	(3.72)	(3.63)	(3.74)	(2.70)	(2.93)	(3.02)	(2.92)	(3.16)
p-value (Generalist Team = Mixed Team)	0.62	0.63	0.96	0.83	0.59	0.61	0.59	0.30
p-value (Generalist Team = Specialist Team)	0.61	0.37	0.50	0.16	0.07	0.12	0.04	0.06
p-value (Specialist Team = Mixed Team)	0.92	0.56	0.48	0.16	0.19	0.16	0.07	0.29
R^2	0.53							
Observations	1,285,813							

Notes: The table reports the estimates of equation (1.10). The dependent variable is the percentile rank of a user's submission among all submissions to a competition c. Each column presents the percentile change in submission rank percentile relative to those of solo competitors for a time period relative to the period immediately before team formation for each team group. All regressions include user × team fixed effects and competition day fixed effects. Standard errors are clustered at the team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

1.C.2 Additional Results for Section 1.4.5

Table 1.C.4: Robustness to Alternative Ability Proxies: Performance Lags

Ability Control:	Lag-1 Solo Performance			Lag-1&2 Solo Performance			Lag-1, 2& 3 Solo Performance		
Dependent Variable:	*Rank*	*Medal Win*	*Top 3*	*Rank*	*Medal Win*	*Top 3*	*Rank*	*Medal Win*	*Top 3*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Generalist Team	20.77***	0.32***	0.06***	21.22***	0.32***	0.06***	20.32***	0.32***	0.06***
	(1.15)	(0.03)	(0.01)	(1.13)	(0.03)	(0.01)	(1.15)	(0.03)	(0.01)
Mixed Team	17.11***	0.26***	0.02**	16.91***	0.26***	0.02***	17.03***	0.27***	0.02***
	(0.93)	(0.02)	(0.01)	(0.93)	(0.02)	(0.01)	(0.94)	(0.02)	(0.01)
Specialist Team	15.63***	0.23***	0.01	15.65***	0.23***	0.01	16.39***	0.26***	0.01
	(1.79)	(0.03)	(0.01)	(1.99)	(0.04)	(0.01)	(2.18)	(0.04)	(0.01)
3-Person Team	6.73***	0.18***	0.04**	6.69***	0.19***	0.04**	6.55***	0.19***	0.04**
	(1.60)	(0.04)	(0.02)	(1.43)	(0.04)	(0.02)	(1.45)	(0.04)	(0.02)
4-Person Team	12.87***	0.30***	0.05	11.14***	0.28***	0.05	10.27***	0.26***	0.05
	(2.13)	(0.06)	(0.03)	(2.17)	(0.06)	(0.03)	(2.19)	(0.06)	(0.03)
5-Person Team	16.82***	0.30***	-0.00	14.88***	0.28***	-0.01	13.28***	0.28***	-0.01
	(1.80)	(0.06)	(0.02)	(1.80)	(0.06)	(0.02)	(1.85)	(0.06)	(0.02)
R^2	0.17	0.12	0.05	0.19	0.14	0.05	0.21	0.16	0.06
Observations	60090	60090	60090	51732	51732	51732	45557	45557	45557
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.01	0.09	0.00	0.00	0.10	0.01	0.02	0.18	0.01
p-value (Generalist Team = Specialist Team)	0.01	0.04	0.00	0.01	0.06	0.00	0.10	0.23	0.01
p-value (Specialist Team = Mixed Team)	0.44	0.39	0.51	0.55	0.45	0.57	0.78	0.78	0.72

Notes: The table reports the estimates of equation (1.8) with alternative ability proxies. The dependent variables are the percentile ranks of a team's solution, an indicator whether a team's solution has won a medal, and an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. All regressions include competition experience, measuring the average of the number of competitions each team member participated in before competition c. Ability controls are the average percentile rank achieved by each team member in their last, second to last, and third to last solo competition (lag 1, 2 and 3 solo rank) for dependent variables *Rank* and *Top 3*, or the share of team members who won a medal in their last, second to last, and third to last solo competition (lag 1, 2 and 3 solo rank). Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.C.5: Robustness to Alternative Ability Proxies: Performance Change

Ability Control:	L1-L2 Solo Rank Change		L2-L3 Solo Rank Change		Average Solo Rank Change	
Dependent Variable:	*Rank*	*Top 3*	*Rank*	*Top 3*	*Rank*	*Top 3*
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	22.78***	0.06***	22.89***	0.06***	22.88***	0.06***
	(1.21)	(0.01)	(1.25)	(0.01)	(1.25)	(0.01)
Mixed Team	18.52***	0.02***	18.73***	0.02***	18.72***	0.02***
	(0.99)	(0.01)	(1.02)	(0.01)	(1.02)	(0.01)
Specialist Team	16.09***	0.01	16.84***	0.01	16.82***	0.01
	(1.95)	(0.01)	(2.03)	(0.01)	(2.03)	(0.01)
3-Person Team	6.97***	0.04**	6.75***	0.04**	6.77***	0.04**
	(1.49)	(0.02)	(1.53)	(0.02)	(1.53)	(0.02)
4-Person Team	12.30***	0.05	11.36***	0.05	11.37***	0.05
	(2.08)	(0.03)	(2.12)	(0.03)	(2.12)	(0.03)
5-Person Team	15.69***	-0.01	14.80***	-0.01	14.83***	-0.01
	(1.79)	(0.02)	(1.86)	(0.02)	(1.86)	(0.02)
R^2	0.11	0.05	0.11	0.06	0.11	0.06
Observations	51732	51732	45557	45557	45557	45557
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.00	0.01	0.00	0.01	0.00	0.01
p-value (Generalist Team = Specialist Team)	0.00	0.00	0.01	0.01	0.01	0.01
p-value (Specialist Team = Mixed Team)	0.24	0.56	0.38	0.70	0.38	0.70

Notes: The table reports the estimates of equation (1.8) with alternative ability proxies. The dependent variables are the percentile rank of a team's solution and an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. All regressions include competition experience, measuring the average of the number of competitions each team member participated in before competition c. Ability controls is the team-level average of the change in solo ranks between the last and second to last, between second to last and third to last, or their average. Demographic controls include share female, average user platform age, country shares for all countries with at least 100 users, and occupation shares for all occupations reported by at least 100 users (c.f. Table 1.B.2). Code quality controls include the average of each team member's lines of code, lines of comment, number of code files, and votes received by code files. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.C.6: Robustness to Different Specialization Indices

Specialization Index:	Excluding Post-ChatGPT Code			Python Package Diversity		
Dependent Variable:	*Rank*	*Medal Win*	*Top 3*	*Rank*	*Medal Win*	*Top 3*
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	19.50***	0.32***	0.05***	18.12***	0.27***	0.03***
	(1.25)	(0.03)	(0.01)	(0.82)	(0.02)	(0.01)
Mixed Team	16.72***	0.26***	0.02***	17.03***	0.27***	0.03***
	(0.91)	(0.02)	(0.01)	(1.18)	(0.02)	(0.01)
Specialist Team	18.95***	0.24***	0.01	9.47**	0.15**	0.02
	(1.72)	(0.03)	(0.01)	(4.07)	(0.07)	(0.02)
Competition Experience	0.08***	-0.00***	-0.00***	0.08***	-0.00***	-0.00***
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Previous Rank	0.50***		0.00***	0.49***		0.00***
	(0.01)		(0.00)	(0.01)		(0.00)
Previous Medal Wins		0.39***			0.39***	
		(0.01)			(0.01)	
3-Person Team	6.07***	0.18***	0.04**	5.91***	0.17***	0.04**
	(1.64)	(0.04)	(0.02)	(1.62)	(0.04)	(0.02)
4-Person Team	12.36***	0.29***	0.05	11.51***	0.28***	0.04
	(2.21)	(0.06)	(0.03)	(2.17)	(0.06)	(0.03)
5-Person Team	14.55***	0.29***	-0.01	13.64***	0.27***	-0.02
	(1.80)	(0.06)	(0.02)	(1.75)	(0.06)	(0.02)
R^2	0.21	0.16	0.05	0.20	0.15	0.05
Observations	56183	56183	56183	59047	59047	59047
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.05	0.06	0.07	0.39	1.00	0.94
p-value (Generalist Team = Specialist Team)	0.79	0.06	0.03	0.04	0.09	0.57
p-value (Specialist Team = Mixed Team)	0.23	0.63	0.44	0.07	0.10	0.56

Notes: The table reports the estimates of equation (1.8) with alternative ways of classifying specialists and generalists. Columns 1-3 use the specialization index as described in Section 1.3.3, excluding all code files published after the introduction of ChatGPT in November 2022 from the code base. Columns 4-6 use a measure of code diversity based on used python packages (c.f. section 1.4.5). The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. Standard errors are clustered at the stable team level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

1.D Additional Results for Section 1.5

1.D.1 Additional Results for Section 1.5.1

Table 1.D.7: Team Type, Complexity, and Solution Quality

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Complexity:	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	20.49***	18.34***	0.33***	0.25***	0.06***	0.05**
	(1.20)	(2.16)	(0.03)	(0.05)	(0.02)	(0.02)
Mixed Team	17.19***	14.52***	0.27***	0.20***	0.01*	0.02*
	(1.03)	(1.96)	(0.02)	(0.03)	(0.01)	(0.01)
Specialist Team	18.59***	8.52**	0.27***	0.09*	0.02	-0.00
	(2.13)	(3.76)	(0.04)	(0.05)	(0.01)	(0.00)
Competition Experience	0.10***	0.03**	-0.00***	-0.00***	-0.00***	-0.00***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.55***	0.34***			0.00***	0.00***
	(0.01)	(0.01)			(0.00)	(0.00)
Previous Medal Wins			0.42***	0.29***		
			(0.01)	(0.02)		
3-Person Team	5.81***	6.99**	0.20***	0.11*	0.06**	0.01
	(1.67)	(3.08)	(0.04)	(0.06)	(0.02)	(0.03)
4-Person Team	13.04***	7.82	0.32***	0.13	0.06*	-0.02**
	(1.62)	(9.75)	(0.06)	(0.15)	(0.04)	(0.01)
5-Person Team	14.30***	16.71***	0.34***	0.23**	-0.02***	0.02
	(2.04)	(3.39)	(0.06)	(0.11)	(0.01)	(0.04)
R^2	0.25	0.11	0.16	0.13	0.06	0.04
Observations	43684	16406	43684	16406	43684	16406
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.02	0.16	0.08	0.38	0.01	0.20
p-value (Generalist Team = Specialist Team)	0.42	0.02	0.18	0.02	0.02	0.01
p-value (Specialist Team = Mixed Team)	0.54	0.15	0.97	0.06	0.86	0.02

Notes: The table reports the estimates of equation (1.8), separately for the sample of low and high complexity competitions. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.D.8: Robustness to Complexity Measure: Cutoff at 50$^{\text{th}}$ percentile

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Complexity (Cutoff at 50$^{\text{th}}$ percentile):	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	21.24***	19.14***	0.36***	0.27***	0.08***	0.04***
	(1.50)	(1.42)	(0.04)	(0.03)	(0.02)	(0.01)
Mixed Team	17.80***	15.75***	0.29***	0.23***	0.01	0.02***
	(1.34)	(1.22)	(0.03)	(0.02)	(0.01)	(0.01)
Specialist Team	20.17***	13.58***	0.29***	0.18***	0.00	0.01
	(2.59)	(2.44)	(0.05)	(0.04)	(0.01)	(0.01)
Competition Experience	0.08***	0.07***	-0.00***	-0.00**	-0.00***	-0.00***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.56***	0.45***			0.00***	0.00***
	(0.01)	(0.01)			(0.00)	(0.00)
Previous Medal Wins			0.45***	0.33***		
			(0.02)	(0.01)		
3-Person Team	5.84***	6.04***	0.19***	0.16***	0.07**	0.02
	(1.83)	(2.32)	(0.05)	(0.05)	(0.03)	(0.02)
4-Person Team	11.52***	13.45***	0.28***	0.31***	0.02	0.07
	(2.31)	(3.32)	(0.08)	(0.08)	(0.04)	(0.05)
5-Person Team	16.37***	13.75***	0.45***	0.19**	-0.02*	0.00
	(1.95)	(2.53)	(0.06)	(0.08)	(0.01)	(0.03)
R^2	0.25	0.17	0.19	0.13	0.06	0.04
Observations	25942	34148	25942	34148	25942	34148
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.06	0.05	0.09	0.28	0.01	0.13
p-value (Generalist Team = Specialist Team)	0.72	0.04	0.27	0.07	0.00	0.10
p-value (Specialist Team = Mixed Team)	0.40	0.41	0.97	0.30	0.62	0.65

Notes: The table reports the estimates of equation (1.8), separately for the sample of low and high complexity competitions. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.D.9: Robustness to Complexity Measure: Cutoff at 90$^{\text{th}}$ percentile

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Complexity (Cutoff at 90$^{\text{th}}$ percentile):	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	19.74***	21.01***	0.30***	0.30***	0.06***	0.04
	(1.12)	(3.42)	(0.03)	(0.07)	(0.01)	(0.03)
Mixed Team	17.27***	10.73***	0.27***	0.11**	0.01**	0.03
	(0.93)	(3.52)	(0.02)	(0.05)	(0.01)	(0.02)
Specialist Team	17.31***	5.21	0.25***	-0.05	0.01	0.00
	(2.11)	(5.47)	(0.03)	(0.05)	(0.01)	(0.00)
Competition Experience	0.09***	-0.02	-0.00***	-0.00**	-0.00***	-0.00***
	(0.01)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.53***	0.25***			0.00***	0.00***
	(0.01)	(0.02)			(0.00)	(0.00)
Previous Medal Wins			0.40***	0.30***		
			(0.01)	(0.03)		
3-Person Team	6.69***	1.88	0.19***	0.05	0.05**	-0.04**
	(1.50)	(5.55)	(0.04)	(0.09)	(0.02)	(0.02)
4-Person Team	12.47***	9.36	0.30***	0.22	0.05	-0.03
	(1.90)	(15.12)	(0.06)	(0.20)	(0.03)	(0.02)
5-Person Team	13.86***	20.54**	0.28***	0.32	-0.00	-0.04*
	(1.74)	(8.90)	(0.06)	(0.21)	(0.02)	(0.02)
R^2	0.23	0.08	0.16	0.13	0.05	0.03
Observations	52540	7550	52540	7550	52540	7550
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.06	0.03	0.27	0.03	0.00	0.71
p-value (Generalist Team = Specialist Team)	0.30	0.01	0.23	0.00	0.00	0.16
p-value (Specialist Team = Mixed Team)	0.99	0.38	0.68	0.02	0.76	0.13

Notes: The table reports the estimates of equation (1.8), separately for the sample of low and high complexity competitions. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.D.10: Robustness to Alternative Complexity Measure: Competition Difficulty

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Difficulty:	Low	High	Low	High	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	19.51***	21.24***	0.32***	0.27***	0.07***	0.04*
	(1.26)	(2.09)	(0.03)	(0.05)	(0.02)	(0.02)
Mixed Team	17.62***	13.49***	0.27***	0.19***	0.01*	0.03**
	(1.03)	(2.00)	(0.02)	(0.04)	(0.01)	(0.01)
Specialist Team	17.67***	10.41***	0.25***	0.12**	0.01	-0.01
	(2.19)	(3.59)	(0.04)	(0.05)	(0.01)	(0.00)
Competition Experience	0.09***	0.04***	-0.00***	-0.00***	-0.00***	-0.00***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.53***	0.39***			0.00***	0.00***
	(0.01)	(0.01)			(0.00)	(0.00)
Previous Medal Wins			0.42***	0.31***		
			(0.01)	(0.02)		
3-Person Team	5.68***	6.58**	0.18***	0.15**	0.05**	0.02
	(1.62)	(3.30)	(0.04)	(0.07)	(0.02)	(0.03)
4-Person Team	11.58***	14.29**	0.28***	0.34***	0.05	0.06
	(2.24)	(6.38)	(0.06)	(0.12)	(0.03)	(0.08)
5-Person Team	16.22***	12.57***	0.35***	0.19	0.00	-0.03**
	(1.49)	(4.18)	(0.06)	(0.12)	(0.02)	(0.01)
R^2	0.23	0.14	0.16	0.12	0.05	0.04
Observations	44944	15146	44944	15146	44944	15146
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.21	0.00	0.15	0.18	0.00	0.60
p-value (Generalist Team = Specialist Team)	0.46	0.01	0.14	0.04	0.01	0.03
p-value (Specialist Team = Mixed Team)	0.98	0.45	0.63	0.27	0.87	0.01

Notes: The table reports the estimates of equation (1.8), separately for the sample of low and high complexity competitions. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.D.11: Robustness to Alternative Complexity Measure: Instruction Length

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Instruction Length:	Short	Long	Short	Long	Short	Long
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	20.92***	19.53***	0.33***	0.29***	0.06***	0.06***
	(1.78)	(1.37)	(0.04)	(0.03)	(0.02)	(0.02)
Mixed Team	17.84***	16.17***	0.29***	0.23***	0.01	0.02**
	(1.49)	(1.13)	(0.03)	(0.02)	(0.01)	(0.01)
Specialist Team	21.90***	13.41***	0.33***	0.17***	-0.00	0.01
	(2.61)	(2.27)	(0.06)	(0.04)	(0.00)	(0.01)
Competition Experience	0.10***	0.07***	-0.00	-0.00***	-0.00***	-0.00***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.50***	0.49***			0.00***	0.00***
	(0.01)	(0.01)			(0.00)	(0.00)
Previous Medal Wins			0.39***	0.38***		
			(0.02)	(0.01)		
3-Person Team	2.54	7.54***	0.14**	0.20***	0.04	0.04*
	(3.00)	(1.70)	(0.06)	(0.04)	(0.03)	(0.02)
4-Person Team	9.48***	13.52***	0.24**	0.32***	0.06	0.05
	(2.58)	(2.61)	(0.12)	(0.06)	(0.07)	(0.03)
5-Person Team	13.85***	15.53***	0.41***	0.29***	-0.02	-0.00
	(3.49)	(1.97)	(0.14)	(0.06)	(0.01)	(0.02)
R^2	0.23	0.18	0.16	0.14	0.06	0.04
Observations	24558	35532	24558	35532	24558	35532
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.16	0.04	0.47	0.11	0.05	0.03
p-value (Generalist Team = Specialist Team)	0.75	0.02	0.94	0.01	0.00	0.03
p-value (Specialist Team = Mixed Team)	0.17	0.25	0.54	0.14	0.16	0.83

Notes: The table reports the estimates of equation (1.8), separately for the sample of competitions with short (below median) and long instructions (above median instruction length). Competition instructions are extracted from the overview page of each competition on Kaggle (competition description). The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table 1.D.12: Robustness to Alternative Complexity Measure: Compute Constraints

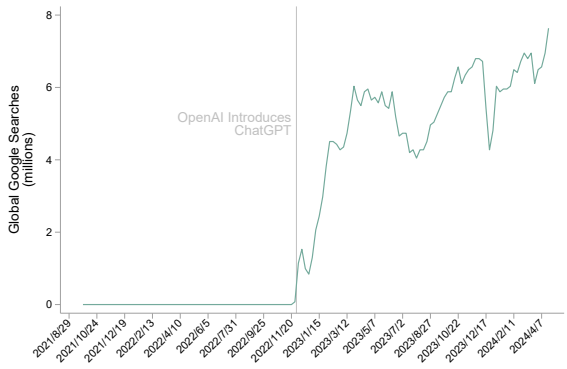
Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
Compute Constraints:	No	Yes	No	Yes	No	Yes
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	20.15***	19.93***	0.32***	0.29***	0.05***	0.07***
	(1.39)	(1.58)	(0.03)	(0.03)	(0.02)	(0.02)
Mixed Team	14.27***	18.41***	0.24***	0.26***	0.01	0.02***
	(1.31)	(1.18)	(0.03)	(0.03)	(0.01)	(0.01)
Specialist Team	19.84***	12.68***	0.29***	0.16***	0.02	0.00
	(2.10)	(2.80)	(0.05)	(0.04)	(0.02)	(0.01)
Competition Experience	0.07***	0.09***	-0.00**	-0.00***	-0.00***	-0.00***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.51***	0.47***			0.00***	0.00***
	(0.01)	(0.01)			(0.00)	(0.00)
Previous Medal Wins			0.40***	0.37***		
			(0.01)	(0.02)		
3-Person Team	3.68*	7.56***	0.15***	0.20***	0.07**	0.03
	(2.19)	(1.94)	(0.05)	(0.04)	(0.03)	(0.02)
4-Person Team	9.06**	13.73***	0.26***	0.31***	0.03	0.06
	(4.16)	(2.35)	(0.10)	(0.07)	(0.04)	(0.04)
5-Person Team	12.07***	15.21***	0.27**	0.30***	-0.02**	-0.00
	(3.82)	(1.85)	(0.10)	(0.07)	(0.01)	(0.02)
R^2	0.23	0.17	0.17	0.13	0.06	0.04
Observations	31588	28502	31588	28502	31588	28502
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.00	0.40	0.07	0.39	0.09	0.01
p-value (Generalist Team = Specialist Team)	0.90	0.02	0.68	0.01	0.29	0.00
p-value (Specialist Team = Mixed Team)	0.02	0.05	0.35	0.03	0.57	0.06

Notes: The table reports the estimates of equation (1.8), separately for the sample of low and high complexity competitions, where high complexity competitions are those with compute constraints. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

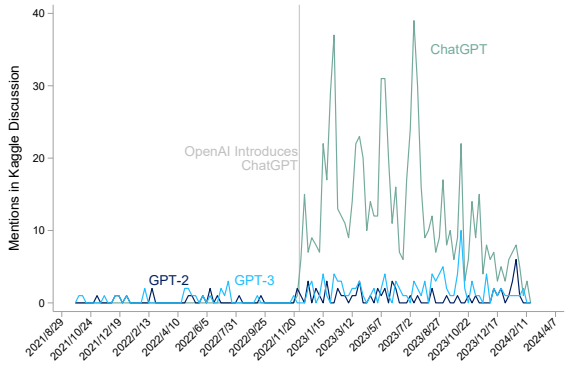
1.D.2 Additional Results for Section 1.5.2

Figure 1.D.4: Google Searches and Kaggle Forum Mentions of ChatGPT

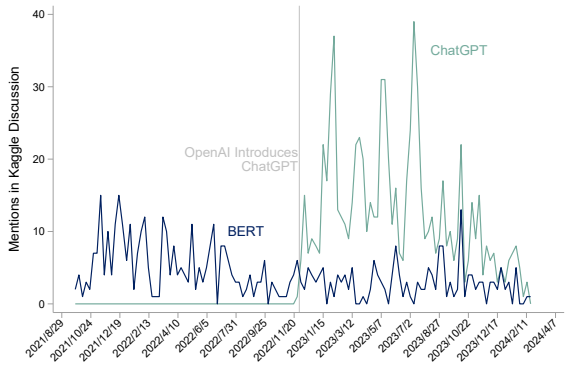
(a) Google Searches of ChatGPT



(b) Forum Mentions of ChatGPT, GPT-2, and GPT-3



(c) Forum Mentions of ChatGPT and BERT



Notes: Panel (a) plots google searches of the term "ChatGPT" for the time period between September 2021 and July 2024. Data come from Google Trends, converted to search counts using Glimpse-Google Trends Supercharged Plugin. Panel (b) plots mentions of the term "ChatGPT" compared to "GPT-2" and "GPT-3", and Panel (c) compared to "BERT" in public Kaggle forums for the time period between September 2021 and July 2024.

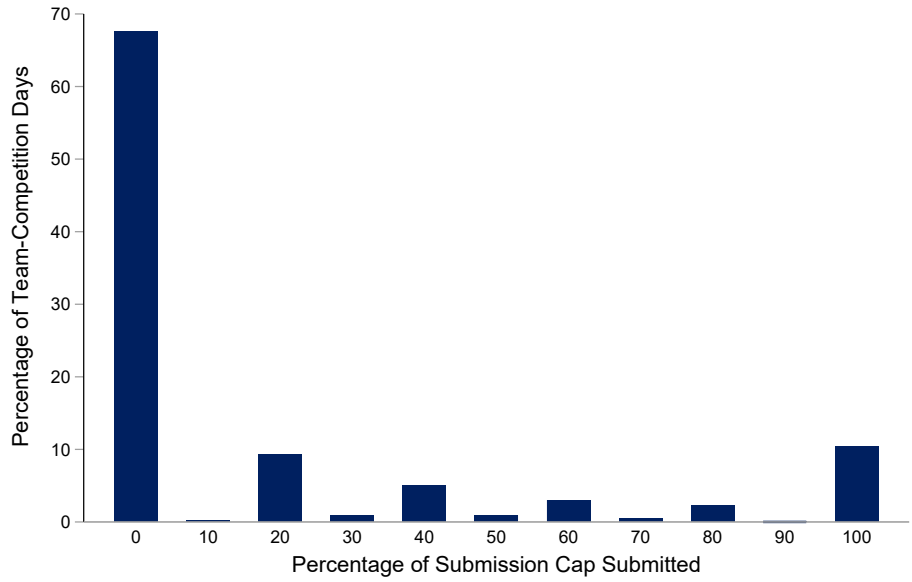
Table 1.D.13: Team Type, ChatGPT, and Solution Quality

Dependent Variable:	*Rank*		*Medal Win*		*Top 3*	
ChatGPT:	Pre	Post	Pre	Post	Pre	Post
	(1)	(2)	(3)	(4)	(5)	(6)
Generalist Team	21.60***	15.26***	0.34***	0.34***	0.04*	0.05
	(2.45)	(3.13)	(0.06)	(0.06)	(0.02)	(0.04)
Mixed Team	18.11***	14.72***	0.28***	0.22***	0.03*	0.01
	(2.50)	(2.38)	(0.05)	(0.05)	(0.02)	(0.02)
Specialist Team	10.09***	17.26***	0.05	0.24***	0.02	-0.01***
	(3.91)	(3.99)	(0.07)	(0.06)	(0.03)	(0.00)
Competition Experience	0.09***	0.05***	-0.00	-0.00***	-0.00***	-0.00***
	(0.01)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
Previous Rank	0.51***	0.48***			0.00*	0.00
	(0.02)	(0.02)			(0.00)	(0.00)
Previous Medal Wins			0.38***	0.37***		
			(0.03)	(0.03)		
3-Person Team	2.50	12.86***	0.10	0.29***	0.06	0.04
	(3.35)	(3.65)	(0.07)	(0.08)	(0.04)	(0.04)
4-Person Team	17.23***	15.78***	0.40***	0.25*	0.07	0.08
	(2.98)	(4.95)	(0.10)	(0.14)	(0.11)	(0.11)
5-Person Team	17.25***	15.56***	0.31***	0.42***	-0.03**	-0.02*
	(3.54)	(5.98)	(0.10)	(0.15)	(0.02)	(0.01)
R^2	0.20	0.16	0.15	0.11	0.05	0.04
Observations	8093	9946	8093	9946	8093	9946
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Code Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes
Competition FEs	Yes	Yes	Yes	Yes	Yes	Yes
p-value (Generalist Team = Mixed Team)	0.29	0.88	0.42	0.11	0.76	0.35
p-value (Generalist Team = Specialist Team)	0.01	0.69	0.00	0.29	0.51	0.10
p-value (Specialist Team = Mixed Team)	0.06	0.57	0.01	0.76	0.64	0.14

Notes: The table reports the estimates of equation (1.8), separately for the sample of competitions before and after the introduction of ChatGPT. The dependent variable is the percentile rank of a team's solution quality, an indicator whether a team's solution has won a medal, or an indicator whether a team's solution was among the top three solutions in competition c. The main explanatory variables are indicators for the team's type: Generalist Team (a team of only generalists), Mixed Team (a team of both generalists and specialists), and Specialist Team (a team of only specialists). Solo competitors are the omitted category. Standard errors are clustered at the level of the stable team. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

1.D.3 Additional Results for Section 1.5.3

Figure 1.D.5: Percentage of Daily Submission Caps



Notes: The figure plots how frequently submission caps are exhausted to which extent. Each category on the x-axis corresponds to a share of allowed submissions, i.e., 0 indicates that no submissions were made on a given day, and 100 that all possible submissions were made on a given day. The y-axis shows the share of team-competition days a submission share was reached. Only on circa 10% of days, teams reach the submission cap. Note that this includes competitions in which the daily submission cap is one submission.

Chapter 2

Climbing the Ivory Tower: How Socioeconomic Background Shapes Academia

---

with Ran Abramitzky, Santiago Pérez, Joseph Price, Carlo Schwarz and Fabian Waldinger

This chapter is based on Abramitzky et al. (2024a). I gratefully acknowledge financial support from the German Research Foundation (DFG) through CRC-TRR 190.

2.1 Introduction

The underrepresentation of individuals from lower socio-economic backgrounds in leadership positions in government, business, and academia has become a growing concern among policymakers and the general public. Efforts to increase representation are driven by two primary economic rationales. First, disparities in the representation of societal groups raise concerns regarding fairness and equality of opportunity. Second, unequal representation can undermine efficiency, as the misallocation of talent deprives society of valuable contributions from individuals in underrepresented groups (Hsieh et al., 2019). In knowledge creation sectors, such as academia, this underrepresentation introduces an additional inefficiency: the unique lived experiences of underrepresented groups offer valuable perspectives that could diversify and enrich the scope of ideas that are explored (e.g., Thorp, 2023). In essence, the absence of these individuals – *missing people* – can lead to *missing ideas*, which is particularly problematic in a world where ideas may be "getting harder to find" (Bloom et al., 2020).

In this paper, we explore how socio-economic background shapes academia – from who becomes an academic through the research fields professors specialize in, to their productivity and peer recognition. For our analysis, we assemble the most comprehensive data on the socio-economic backgrounds and research output of U.S. academics. The long-run nature and granularity of our data enable us to study how these patterns changed over time and how they differ by discipline and across universities.

We rely on three primary data sources to assemble our data. First, we utilize comprehensive faculty rosters from the *World of Academia Database* (Iaria et al., 2024), which provides detailed information on the name, discipline, and academic rank of nearly all academics at U.S. universities from 1900 to 1969. A key advantage of these data is that they list academics regardless of whether they publish or whether they are members of academic societies. This helps to mitigate selection biases common in studies that rely exclusively on publication databases, surveys, or lists of distinguished scholars. Second, we measure the socio-economic background of academics by linking these faculty rosters to full-count U.S. censuses. We then link academics to their family backgrounds using data from the *Census Linking Project* (CLP) (Abramitzky et al., 2021a) and the *Census Tree Project* (Buckles et al., 2023). Our measure of socio-economic background is the percentile rank of their father's predicted income when the future academics were growing up.[1] Third, we link academics in six scientific disciplines – medicine, biology, biochemistry, chemistry, physics, and mathematics – to their publication and citation records using data from the *Clarivate Web of Science*. Overall, our data enable us to measure the socio-economic backgrounds of 46,139 academics (for 15,521 of whom we also have publication and citation data) across 1,026 universities

[1]The findings are robust to alternative measures of socio-economic background.

over nearly seven decades.

Our paper is organized into four parts, examining key stages of academic careers and how they are shaped by socioeconomic background. In the first part of the paper, we examine differential barriers to entry into academia. We find a stark underrepresentation of individuals from lower socio-economic backgrounds: those born to parents in the bottom quintile of the parental income distribution account for less than 5% of all academics. In contrast, around half of U.S. academics come from the top quintile of the income rank distribution. Children born to the highest-earning fathers are particularly overrepresented, with those born to fathers in the 100^{th} percentile having a 56% higher chance of becoming an academic than those born to fathers in the 99^{th} percentile. The underrepresentation of low socio-economic status individuals in academia is greater than in other occupations that require specialized training, such as medicine and law.

We find that the socio-economic composition of academics has remained remarkably stable over seven decades, despite significant changes in American higher education and society – including a sharp increase in college attendance rates. This persistence stands in stark contrast with the significant increase in the representation of women in U.S. academia over the same period (e.g., Rossiter, 1982, 1998; Iaria et al., 2024).

While academics from low socio-economic backgrounds are underrepresented in all universities, the underrepresentation of academics from low socio-economic backgrounds varies sharply by university. In selective private universities such as Princeton, Harvard, and Yale, at least 60% of academics come from families in the top quintile of the parental income distribution. In contrast "only" 30-40% of academics in state universities such as Iowa State, University of Missouri, or the University of Nebraska come from families in this quintile.

Representation also varies sharply by discipline. While around 60% of academics in the humanities come from the top quintile of the parental income distribution, around 40% of academics in mathematics and economics come from the top quintile. This heterogeneity appears to be systematically related to the types of skills required to enter a discipline. Specifically, we find that representation from lower socio-economic backgrounds is higher in disciplines with a stronger emphasis on quantitative relative to verbal skills.

In the second part of the paper, we study the extent to which the influence of parental *occupation* can explain differences in representation by discipline. We develop a novel measure of overrepresentation to assess whether children of fathers in specific occupations are overrepresented in particular academic disciplines. Our findings indicate that academics tend to pursue disciplines aligned with their fathers' occupations. For example, the children of architects are more likely to become professors in architecture, children of artists are more likely to become professors of arts and design, children of bank tellers are more likely to become professors in business and management,

and children of lawyers are more likely to become professors in law. Additionally, using a text embeddings model, we determine the semantic proximity of a father's occupation (e.g., "farmer") to an academic discipline (e.g., "agriculture"). This allows us to identify the discipline that is closest in semantic space to the father's occupation. We then show that academics are more likely to enter disciplines that are systematically similar to their fathers' occupations. Overall, these findings indicate that socio-economic background affects not only the probability of becoming an academic but also the specific discipline that academics pursue.

In the third part of the paper, we study how socio-economic background relates to scholars' productivity. We find no systematic relationship between parental income ranks and the *average* number of publications of academics. However, individuals from lower socio-economic backgrounds are both significantly more likely to never publish and more likely to have a publication count in the top 1%.

Importantly, academics from lower socio-economic backgrounds differ in the *content* of their research. To examine potential differences in a key dimension of publication content, we develop a metric that captures the number of novel words that a scientist introduced to the scientific community (Iaria et al., 2018). The measure proxies for the introduction of new scientific concepts that required novel scientific terms. We find that scientists with a low-income father (father at the 25th percentile) publish around 0.05 additional papers (or 17% more papers) with at least one novel word compared to scientists whose fathers were at the 75th percentile. These findings suggest that academics from lower socio-economic backgrounds are more likely to pursue research agendas off the beaten path, which may result in scientific breakthroughs but also in a higher failure rate, making them riskier hires.

In the fourth part of the paper, we examine the relationship between socio-economic background and recognition by other academics. We start by studying citations to academic papers, a widely used metric for measuring recognition within the academic community. We find that papers published by authors from lower socio-economic backgrounds receive fewer citations. To further explore how socio-economic background affects recognition, we investigate Nobel Prize nominations and awards – an acknowledgment for exceptional scientific contributions. We find that scientists whose fathers were at the 75th percentile of the income rank are around 0.6 percentage points (or 50%) more likely to be nominated for a Nobel Prize than scientists with fathers at the 25th percentile. They are also 50% more likely to be awarded a Nobel Prize. These differences persist even if we control for scientists' publication and citation records.

Our paper contributes to a fast-growing literature on the backgrounds of high-skilled, "elite" professionals such as politicians (Dal Bó et al., 2017) or civil servants (Moreira and Pérez, 2022). It is particularly close to research documenting the socio-economic background of inventors (Bell et al.,

2019; Aghion et al., 2018, 2023; Akcigit et al., 2017) and concurrent research on academics (Morgan et al., 2022; Airoldi and Moser, 2024; Stansbury and Schultz, 2023; Stansbury and Rodriguez, 2024; Novosad et al., 2024).[2] We contribute to this literature with the most comprehensive analysis of the socio-economic background of U.S. academics covering all disciplines and the near universe of universities. The time dimension of our data allows us to trace the evolution of the socio-economic background over a key period in the history of U.S. higher education from the "formative" prewar years, to the consolidation of American leadership in higher education after World War II. The granular nature of our data enables us to advance the literature by studying how hiring, discipline choice, productivity, and recognition are shaped by the socio-economic background of academics. Other related research has documented the importance of socio-economic background for the selection of *students* into elite universities (Chetty et al., 2020; Michelman et al., 2022; Chetty et al., 2023; Abramitzky et al., 2024b).

Our paper is also related to the literature on gender discrimination in academia (e.g., Card et al., 2020, 2022; Iaria et al., 2024; Ross et al., 2022; Moser and Kim, 2022; Koffi, 2024; Hengel, 2022; Babcock et al., 2017; Bagues et al., 2017). While this substantial body of research has studied the underrepresentation of women in research, the underrepresentation of individuals from lower socio-economic backgrounds has been a "forgotten dimension of diversity" (Ingram, 2021), which we examine in this paper.

Finally, we contribute to the literature on how scientists' or inventors' background shapes their research focus and, thereby, the direction of innovation. Existing work by Koning et al. (2021); Einio et al. (2022); Kozlowski et al. (2022); Truffa and Wong (2022); Kozlowski et al. (2022); Dossi (2024); Croix and Goñi (2024) investigates how gender and race impact the research focus of scientists. One of the few papers that studies how socio-economic background affects the direction of research is a recent contribution by Einio et al. (2022). They document that inventors from poorer backgrounds are more likely to patent "necessity" interventions. To the best of our knowledge, we provide the first systematic evidence of how the socio-economic background shapes the research of university academics. Since most basic research, as well as the training of future innovators, occurs in universities, the selection of academics likely has important knock-on effects for downstream innovation.

[2]Similarly, geography also shapes participation in science. Participants of the international mathematical olympiads from lower-income countries are less likely to enroll in PhD programs and produce fewer publications and citations despite similar talents (Agarwal and Gaule, 2020).

2.2 Data

For our analysis, we construct the largest individual-level dataset of U.S. university academics ever assembled, which we combine with information on their socio-economic background and their research output. The dataset is based on three data sources. First, we use complete faculty rosters for the near universe of U.S. universities from the *World of Academia Database* (Iaria et al. 2024). Second, we match these data to historical U.S. censuses (Ruggles et al., 2024). Using links from the *Census Linking Project (CLP)* (Abramitzky et al. 2012, 2021a), the *Census Tree Project* (Buckles et al. 2023) and the *IPUMS Multigenerational Longitudinal Panel (MLP)* (Ruggles et al., 2019) we are able to trace academics to their childhood homes, which enables us to measure the socio-economic background of academics. Third, we enhance the data with publication and citation data from the *Web of Science* to observe the academics' research output and its content.

2.2.1 Historic Faculty Rosters from the World of Academia Database

The *World of Academia Database* contains faculty rosters for nearly all Ph.D.-granting universities in the United States. We use six cross-sections covering U.S. academics in 1900, 1914, 1925, 1938, 1956, and 1969.[3] For example, the data contain 3,441 U.S. academics who entered the database in 1900 and 65,340 U.S. academics who entered the database in 1969, reflecting the spectacular growth of the U.S. university sector during the 20th century (Table 2.2.1).

For the period of our analysis, the database provides the most comprehensive data on academics in the United States (see Iaria et al. 2024 for details and comparisons to other data sources). In addition to academics' names, universities, and academic rank (i.e., assistant, associate, or full professor), we observe their specialization, which we code into 36 disciplines.[4] For example, the 1938 faculty roster lists George Wells Beadle as a Biology professor at Stanford University (Figure 2.2.1, panel a). He received the 1958 Nobel Prize in Physiology/Medicine for the "discovery of the role of genes in biochemical events within cells."

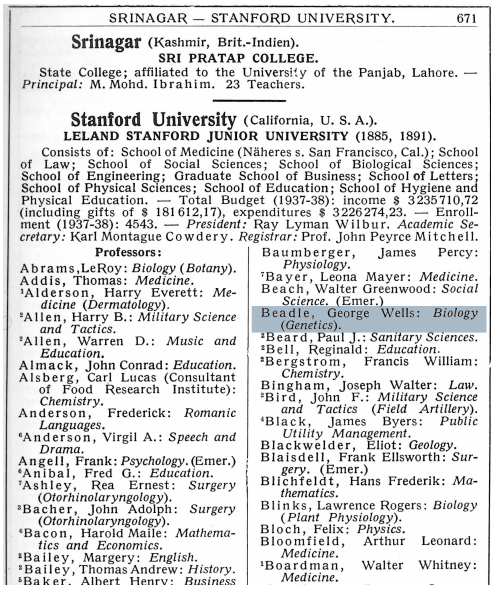
The *World of Academia Database* offers several key features that are integral to our analysis. First, it contains *entire* faculty rosters for the vast majority of PhD granting universities in the United States, which allows us to study academics even if they never published or never became distinguished scientists. This comprehensive coverage enables us to overcome important selection biases that affect studies that rely exclusively on publication or citation databases, surveys, or

[3]The data include all academics who were affiliated with a U.S. university in at least one of the six cross-sections. We thus also include the U.S. spells of academics who start their career abroad and move to the United States or who start their career in the United States and then move abroad. About 10 percent of the academics are only listed with initials in the faculty rosters. As the match to the census data described below uses full first names, we exclude these academics from the data. For the statistics reported in Table 2.2.1, we report their first U.S. cohort in the *World of Academia Database*.

[4]For the vast majority of universities, the data report all academics who are assistant professors and above. Lecturers and similar academic staff are usually not reported.

Figure 2.2.1: Example Data Construction

(a) Sample Page: Faculty Rosters



SRINAGAR — STANFORD UNIVERSITY. 671

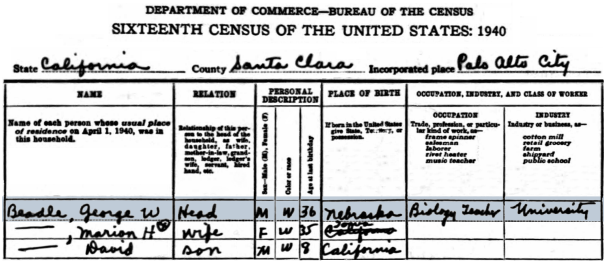
Srinagar (Kashmir, Brit.-Indien).
SRI PRATAP COLLEGE.
State College; affiliated to the University of the Panjab, Lahore. — *Principal:* M. Mohd. Ibrahim. 23 Teachers.

Stanford University (California, U. S. A.).
LELAND STANFORD JUNIOR UNIVERSITY (1885, 1891).
Consists of: School of Medicine (Näheres s. San Francisco, Cal.); School of Law; School of Social Sciences; School of Biological Sciences; School of Engineering; Graduate School of Business; School of Letters; School of Physical Sciences; School of Education; School of Hygiene and Physical Education. — Total Budget (1937-38): income $ 3235710,72 (including gifts of $ 181612,17), expenditures $ 3226274,23. — Enrollment (1937-38): 4543. — *President:* Ray Lyman Wilbur. *Academic Secretary:* Karl Montague Cowdery. *Registrar:* Prof. John Peyrce Mitchell.

Professors:

Abrams, LeRoy: *Biology (Botany).*
Addis, Thomas: *Medicine.*
[1]Alderson, Harry Everett: *Medicine (Dermatology).*
[2]Allen, Harry B.: *Military Science and Tactics.*
[2]Allen, Warren D.: *Music and Education.*
Almack, John Conrad: *Education.*
Alsberg, Carl Lucas (Consultant of Food Research Institute): *Chemistry.*
Anderson, Frederick: *Romanic Languages.*
[6]Anderson, Virgil A.: *Speech and Drama.*
Angell, Frank: *Psychology.* (Emer.)
[6]Anibal, Fred G.: *Education.*
[7]Ashley, Rea Ernest: *Surgery (Otorhinolaryngology).*
[3]Bacher, John Adolph: *Surgery (Otorhinolaryngology).*
[6]Bacon, Harold Maile: *Mathematics and Economics.*
[2]Bailey, Margery: *English.*
[5]Bailey, Thomas Andrew: *History.*
[5]Baker, Albert Henry: *Business*
Baumberger, James Percy: *Physiology.*
[7]Bayer, Leona Mayer: *Medicine.*
Beach, Walter Greenwood: *Social Science.* (Emer.)
Beadle, George Wells: *Biology (Genetics).*
[2]Beard, Paul J.: *Sanitary Sciences.*
[2]Bell, Reginald: *Education.*
[1]Bergstrom, Francis William: *Chemistry.*
Bingham, Joseph Walter: *Law.*
[2]Bird, John F.: *Military Science and Tactics (Field Artillery).*
[4]Black, James Byers: *Public Utility Management.*
Blackwelder, Eliot: *Geology.*
Blaisdell, Frank Ellsworth: *Surgery.* (Emer.)
Blichfeldt, Hans Frederik: *Mathematics.*
Blinks, Lawrence Rogers: *Biology (Plant Physiology).*
Bloch, Felix: *Physics.*
Bloomfield, Arthur Leonard: *Medicine.*
[1]Boardman, Walter Whitney: *Medicine.*

(b) Adult Census

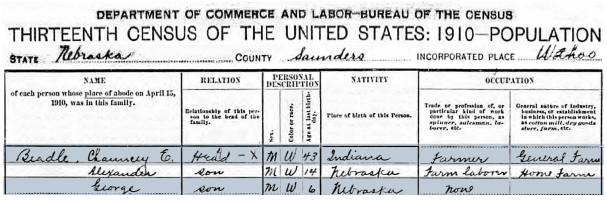


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
SIXTEENTH CENSUS OF THE UNITED STATES: 1940

State California County Santa Clara Incorporated place Palo Alto City

NAME	RELATION	PERSONAL DESCRIPTION: Sex	Color or race	Age at last birthday	PLACE OF BIRTH	OCCUPATION	INDUSTRY
Beadle, George W	Head	M	W	36	Nebraska	Biology Teacher	University
—, Marion H	wife	F	W	35	California		
— David	son	M	W	8	California		

(c) Childhood Census



DEPARTMENT OF COMMERCE AND LABOR—BUREAU OF THE CENSUS
THIRTEENTH CENSUS OF THE UNITED STATES: 1910—POPULATION

State Nebraska County Saunders Incorporated place Wahoo

NAME of each person whose place of abode on April 15, 1910, was in this family.	RELATION: Relationship of this person to the head of the family.	PERSONAL DESCRIPTION: Sex	Color or race	Age at last birthday	NATIVITY: Place of birth of this Person.	OCCUPATION: Trade or profession of, or particular kind of work done by this person, as spinner, salesman, laborer, etc.	OCCUPATION: General nature of industry, business, or establishment in which this person works, as cotton mill, dry goods store, farm, etc.
Beadle, Chauncey C.	Head - X	M	W	43	Indiana	Farmer	General Farm
Alexander	son	M	W	14	Nebraska	Farm Laborer	Home Farm
George	son	M	W	6	Nebraska	None	

Notes: Panel (a) shows a sample page from the faculty roster of Stanford University from the 1938 edition of *Minerva* including the entry of the biology professor "George Wells Beadle." Panel (b) shows George W. Beadle's entry in the 1940 adult census. Panel (c) shows George Beadle's entry in his childhood census (1910) which we use to measure the race, age, state of residence, and occupation of his father ("farmer").

lists of distinguished academics. For instance, lists of distinguished academics might underestimate the number of academics from lower SES-backgrounds if such academics are less likely to be recognized by their peers (as we document below). Second, our dataset encompasses all academic disciplines, including the social sciences and humanities. This broad scope enables us to conduct a comprehensive analysis of representation in academia, examining variations across universities and disciplines.

2.2.2 Measuring Parental Socio-Economic Background

To measure academics' parental socio-economic background, we link the faculty rosters to historical full-count U.S. censuses (Ruggles et al., 2024) using a two-step procedure. In the first step, we link the cross-sections of academics to a contemporaneous U.S. census ("adult census"). In the second step, we use census crosswalks from the Census Linking Project, the Census Tree Project, and IPUMS Multigenerational Longitudinal Panel (MLP) to construct back-links to each academic's childhood census records to measure parental background.

Linking Faculty Rosters to Contemporaneous U.S. Censuses: "Adult Census"

In the first step, we link all academics who appear in the faculty rosters to the two closest contemporaneous censuses. For example, we link the 1925 faculty roster to both the 1920 and 1930 censuses. The only exceptions are the 1956 and 1969 faculty rosters, which can be linked to only one census (the 1950 census) since neither the 1960 nor the 1970 full-count censuses have been released to the public.

We link academics in the faculty rosters to their contemporaneous censuses based on the full name of the academic, their census occupation, and their location in the census.[5] We define a potential match as someone:

1. who has the exact same first and last name in the census and in the faculty rosters
2. whose implied age is between 20 and 100 (based on their age in the census) at the time we observe them in the corresponding faculty rosters
3. who indicates an occupation in the census that aligns with a professorship in a specific discipline (e.g., biology professors may be listed with the occupations "professor", "biologist", or "biology teacher")[6]

We consider all matches that satisfy criteria 1-3 above. If criteria 1-3 only return one potential match between the census and the faculty rosters, we consider the observation pair as matched, and the procedure continues with step 7 (described below). For example, we can link the faculty roster entry of George Wells Beadle to the 1940 census. The unique match in the census reports that he was 36 years old in 1940, lived in Palo Alto City, and worked as a "Biology Teacher" at a "University" (Figure 2.2.1, panel b).

If there are multiple potential matches, we disambiguate them using the following additional criteria:

4. the potential match in the census lives in a county within 150 kilometers of the university reported in the faculty rosters[7]
5. the potential match has the same middle name initial(s) in the census and the faculty rosters

[5] It is important to note, that a relatively small share of professors are listed under the occupation "professor" in the census. Biology professors, for example, are listed as "professor", "biologist", or "biology teacher." This highlights the importance of using faculty rosters to capture university professors instead of using the "professor" occupational category from the census records.

[6] Here, we both use the IPUMS occupation coding (occ1950, see IPUMS (2024a)) as well as the original string responses recorded by the census (occstr). This enables us to also match individuals whose occupation or industry was coded as "not yet classified". Typically, occupations are unclassified due to transcription or spelling errors.

[7] For academics that are affiliated with multiple universities, we calculate the distance between each of their universities and the county and use the minimum distance for disambiguation.

6. the potential match reports an occupation in the census which aligns more closely with their discipline (i.e., if there are two potential matches for a biology professor, one listed in the census as "professor" and the other one as "biology professor," we select the latter observation)

We then keep all matches that are unique after disambiguating them using at least one of the criteria 4-6.

After applying criteria 1–6, approximately 70% of potential matches indicate an industry in the census that aligns with their academic position. For instance, individuals may be listed in industry 888 - Educational Services. Similarly, medical professors are often listed in industry 869 - Hospitals. In contrast, the remaining 30% are listed in industries that do not closely correspond to their academic roles (e.g. 246 - Construction) or fall into an unclassified category. To enhance the reliability of these matches, we introduce a seventh criterion that leverages the specific industry and occupation strings reported in the census:

7. the potential matches must report industry and occupation strings in the census that are consistent with becoming a professor

For the seventh criterion, all potential matches with a misaligned industry are independently reviewed by two research assistants, who classify each link as either correct or incorrect. For instance, the Stanford physics professor Frederick John Rogers was linked to a census record listing the industry as 0 - none reported. The research assistants examined the associated occupation ("Assoct Projessor [sic]") and industry ("physico at Stanford [sic]") strings from the record and determined the match to be correct.[8] In contrast, Vanderbilt University biology professor George W. Martin was linked to a census record listing the industry as 636 - Food stores, except dairy products. The research assistants examined the associated occupation ("druggist") and industry ("own store") strings and classified the link as incorrect. For the analysis, we only retain matches that both research assistants classified as correct.[9]

Throughout the paper, we show results for two different samples:

1. *Main Sample:* 1900-1956 faculty rosters

2. *Extended Sample:* 1900-1969 faculty rosters

We use two different samples because the full-count censuses for 1960 and 1970 are not yet available. It is, therefore, challenging to link individuals who entered the *World of Academia* database in 1969

[8]The misspellings in the occupation and industry fields result from the transcription of handwritten census records.

[9]In cases where we match an academic to multiple census years, we additionally check whether these matches are internally consistent (i.e., that the main demographic information used for backlinking is the same across all matches). For example, an academic matched to a person aged 45 in the 1910 census should match to a person aged 55 in the 1920 census. Our research assistants hand-check all observations for which this is not the case and remove incorrect matches.

Table 2.2.1: Linking Rates

Cohort	Academics entering faculty rosters	Matched to Adult Census		Matched to Childhood Census		
		Total	% Faculty roster	Total	% Adult census	% Faculty roster
Main sample: 1900-1956 cohorts						
1900	3,441	2,485	72.2	1,726	69.5	50.2
1914	5,899	4,487	76.1	3,073	68.5	52.1
1925	6,401	4,731	73.9	3,188	67.4	49.8
1938	23,458	17,792	75.8	12,338	69.3	52.6
1956	53,243	28,814	54.1	17,052	59.2	32.0
Total	92,442	58,309	63.1	37,377	64.1	40.4
Extended sample: 1900-1969 cohorts						
		⋮				
1969	65,340	17,306	26.5	8,762	50.6	13.4
Total	157,782	75,615	47.9	46,139	61.0	29.2

to an adult census. With this in mind, the main sample in our analysis is restricted to academics who we first observe in 1956 or earlier cohorts. However, we also consider an extended sample in which we attempt to match all academics in our data (including those who enter the data in 1969).

Of the 92,442 academics in the main sample, we link 58,309 (63%) to a contemporaneous census (Table 2.2.1).[10] Manual inspections suggest that transcription mistakes of the historical handwritten census records account for many missed links. Furthermore, as we require unique matches based on our linking criteria, we also miss links if matches between the faculty rosters and the census record are not unique. In the extended sample we link 75,615 (48%) to a contemporaneous census (Table 2.2.1). Linking rates are lower for the 1956 and 1969 cohorts for two main reasons. First, these cohorts can only be matched to the 1950 census. Linking to just one adult census lowers the linking rate, as linking to two censuses enables us to deal with idiosyncratic transcription errors occurring in one census but not the other. Second, these cohorts likely include individuals who were not yet academics in 1950 and, hence, cannot be matched on the basis of their census occupation to an adult census.

For each academic that we successfully link to a contemporaneous census, we extract the birth year and the birth state from the adult census. These variables are crucial to link academics to their childhood censuses (see below for more details). For example, we extract George Beadle's birthyear (1903 or 1904, based on the 36 years of age that he reports) and his birth state ("Nebraska") from his 1940 census record (Figure 2.2.1, panel b).

[10] Below, we provide evidence that linked academics are similar to academics who we are unable to link, thereby alleviating selection concerns.

Linking to the Childhood Census to Measure Socio-Economic Background

To construct measures of the socio-economic background of academics, we use census-to-census crosswalks to link the adult census record to the corresponding childhood censuses. First, we use the links available from the Census Linking Project (CLP, Abramitzky et al. 2012, 2021a).[11] We then combine these links with links from the Census Tree Project (CT, Buckles et al. 2023) for the 1900-1940 adult censuses and IPUMS Multigenerational Longitudinal Project (MLP) (Ruggles et al., 2019) for the 1950 adult census.[12] In addition to enabling us to increase the sample size, the additional links allow us to link to the childhood records of some female academics, which are less frequently captured by traditional linking methods.[13]

To maximize the likelihood of capturing an academic's parental background, we link adult census records to all potential childhood censuses. Childhood censuses are defined as those in which future academics are observed as children under the age of 22 and residing with their parents. In cases where an academic is linked to multiple childhood censuses, we prioritize the census in which the academic is youngest.[14]

Our exemplary academic, George Wells Beadle, can be linked to his childhood census of 1910. At the time, he was six years old and listed in the census as the son of Chauncey E. Beadle, who was 43 years old and worked as a farmer (Figure 2.2.1, panel c). The information on the father's occupation will be the key information to reconstruct George Wells Beadle's socio-economic background.

For the main sample, we are able to link 37,377 (or 64% of the adult census) records to a childhood census (Table 2.2.1). For the extended sample, we can link 46,139 (or 61%) of the adult census records to a childhood census.[15] These linking rates are high compared to linking rates in existing research, because we rely on a combination of linking algorithms and since we link to multiple potential childhood censuses.

[11] Specifically, we use the "ABE-exact" links. As of November 2024, the Census Linking Project has not released links between the 1950 census and earlier censuses. Therefore, we create our own crosswalks for the 1950 census using the ABE algorithm in its "exact standard" version.

[12] In the rare cases in which these links point to different individuals, we privilege links made by the ABE exact algorithm. There are few such cases because there is a very high rate of conditional agreement between ABE links and those made by machine learning algorithms, i.e., when both methods identify a link the links are identical in close to 100% of cases (Abramitzky et al., 2021a).

[13] The share of female academics in the faculty rosters is only 13% in the main sample and 14% in the extended sample (see also Iaria et al. 2024). Overall, linking rates for female academics are 28% for the main sample and 21% for the extended sample, compared to 42% and 31% for male academics. All results remain unchanged in a sample of male academics.

[14] As we link some academics to multiple adult censuses that can be linked to different childhood censuses, a small fraction of them have backlinks to different individuals in a childhood census. For example, an individual listed in the 1914 faculty roster could theoretically be matched to both the 1910 and 1920 adult censuses, and the 1920–1880 backlinks might identify a different individual than the 1910–1880 backlinks. In such cases, we retain the backlink associated with the adult census that is closest to the childhood census. In the given example, we would prioritize the link based on the 1910–1880 crosswalk.

[15] For academics who moved to the United States to study or when they were already academics, we cannot link them to a childhood census by construction. Of the 75,615 academics who we link to an adult census, 6,769 or 7.9% are foreign-born. Foreign-born academics are part of the dataset if they migrated as children and can be observed in at least one childhood census after moving to United States.

Overall, we successfully link 37,377 (or 40%) individuals from the main sample to their childhood census. These linked academics form the basis for our analysis. To assess potential selection introduced by our linking procedure, we correlate the department rank (measured as the average number of citations of all academics in a department, see Hager et al. 2024) with the linking rate at the department level. We find no systematic relationship between department quality and the linking rate (Figure 2.2.2, Panels (a) and (b), p-value=0.69).[16] As a further check, we investigate the correlation between the linking rates and the average income associated with a last name.[17] We find no systematic association between these variables (Figure 2.2.2, Panels (c) and (d), p-value=0.36). Together, these results indicate that our linking procedure does not introduce systematic selection.

Constructing Parental SES ranks

For our baseline results, we rely on father's occupational income scores as a proxy for socio-economic background, because other measures of parental socio-economic status such as parental income or parental education are not available in pre-1940 U.S. censuses. We construct parental "income scores" for each academic, following the approach outlined by Abramitzky et al. (2021b). Specifically, we use data on wage income from the 1940 census (the first U.S. census to include individual-level income) and estimate the following regression for all working-age (20-70 years old) men in the 1940 census:

$$\ln(\text{Income}_j) = \beta_0 + \beta_1 \text{Occupation}_j \times \text{State FE} + \beta_2 \text{Age}_j + \beta_3 \text{Age}_j^2 + \beta_4 \text{Race}_j + \epsilon_j \qquad (2.2.1)$$

where $\ln(\text{Income}_j)$ measures the income of individual j in 1940. $\text{Occupation}_j \times \text{State FE}$ is a separate fixed effect for each census occupation code interacted with the state of residence of individual j. In addition, we also include a second-order polynomial in age as well as race fixed effects. Because the 1940 census includes information on income from wages but not on other sources of income, we adjust the income of self-employed farmers using the method developed by Collins and Wanamaker (2022).[18]

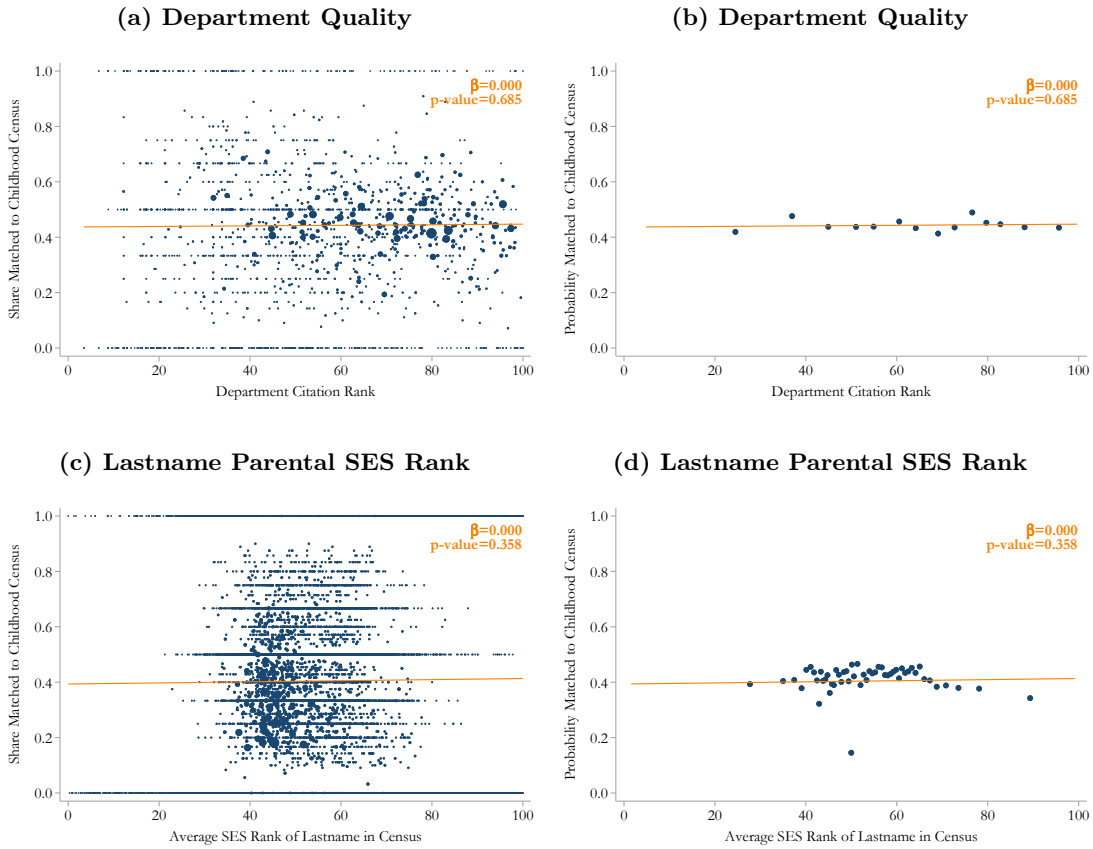
We then use the estimated coefficients from equation (2.2.1) to predict income for fathers in all census years. We use these predicted incomes to rank fathers relative to *all* fathers, including the fathers of non-academics, with children born in the same year. In robustness tests, we construct alternative parental SES ranks based on income predictions that do not differ by state, and also

[16] We report equivalent figures for the extended sample in Figure 2.A.1, Panels (a) and (b). There is a small and marginally significant positive correlation between department quality and matching rates in the extended sample.

[17] We measure the average income of a last name in the census using an analogous procedure to the one described in the next subsection.

[18] In cases where the number of individuals within certain occupation-by-state cells is low, or where census occupation codes change across years (see IPUMS 2024a), we apply coarser fixed effects to predict income ranks. See Appendix 2.A.1 for details.

Figure 2.2.2: Correlation of Linking Rates With Department Quality and Lastname Parental SES Rank



Notes: Panel (a) shows the correlation between a department's citation rank and the probability of linking a scientist to a childhood census for the main sample. Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's parental SES Rank based on the entire U.S. census and the probability of linking an academic to a childhood census. Panel (d) shows a binned scatter plot of the same relationship. Bins are chosen according to Cattaneo et al. (2024). Appendix Figure 2.A.1 shows the equivalent figures for the extended sample.

use alternative measures of socioeconomic status, such as Hisclass (van Leeuwen and Maas, 2011) and Duncan's Socioeconomic Index (SEI).

2.2.3 Linking Scientists with Publications and Citations

To investigate how socio-economic background influences scientific output and the direction of research, we link academics from six scientific disciplines – medicine, biology, biochemistry, chemistry, physics, and mathematics – with publication and citation data from the *Clarivate Web of Science*. We focus on these disciplines for two main reasons. First, they have particularly good coverage in the Web of Science. Second, by the early 20th century, these disciplines had already established

a culture of publishing in scientific journals, with publishing processes resembling contemporary practices. In contrast, disciplines such as the humanities and social sciences predominantly relied on book publishing during this period.

We use the procedure developed by Iaria et al. (2024) to link publications and citations to the faculty rosters. The procedure uses the academic's last name, first name, or initials (depending on whether first names are available), country, city, and discipline.[19] To improve match quality, we harmonize affiliations across the faculty rosters and the *Web of Science* with the *Google Maps API*.

2.2.4 Linking Scientists with Nobel Prize Data

To measure recognition by the scientific community, we hand-link data on nominations for the physics, chemistry, and physiology or medicine Nobel Prizes from the Nobel Nomination archive (Nobelprize.org, 2024). This database contains all nominations for the Nobel Prize in physics and chemistry from 1901 to 1970, and all nominations for the Nobel Prize in physiology or medicine from 1901 to 1953. We also hand-link all Nobel Prize winners to our faculty rosters. Table 2.2.2 provides summary statistics for the most important variables in our data.

Table 2.2.2: Summary Statistics

Panel A: 1900 – 1956			
Variable	Mean	SD	Observations
Parental SES Rank	72.83	24.84	37,377
Age at Entry into Faculty Rosters	45.34	10.11	37,377
Female	0.09		37,377
Publications	4.66	9.51	12,767
Papers with Novel Words	0.30	1.12	11,964
Nominated for Nobel Prize	0.01		12,767
Awarded Nobel Prize	0.00		12,767
Panel B: 1900 – 1969			
Parental SES Rank	72.18	25.06	46,139
Age at Entry into Faculty Rosters	47.28	10.92	46,139
Female	0.10		46,139
Publications	4.91	10.63	15,521
Papers with Novel Words	0.29	1.12	14,718
Nominated for Nobel Prize	0.01		15,521
Awarded Nobel Prize	0.00		15,521

Notes: The table reports summary statistics. Panel A reports information for the main sample, which includes academics who enter the faculty rosters by the 1956 cohort. Panel B reports information for the extended sample, which includes academics who enter the faculty rosters by the 1969 cohort. Data on academics come from the *World of Academia Database*. Parental SES ranks are constructed based on U.S. census microdata. Data on publications come from the *Web of Science*. Publications are measured in a ± 5-year window around the year of entering the faculty rosters. Papers with novel words measures the number of papers published in a ± 5-year window around the year of entering the faculty rosters that introduce at least one novel word. Nominated for Nobel Prize is an indicator whether a scientist was ever nominated for a Nobel Prize, and Awarded Nobel Prize is an indicator for winning the Nobel Prize.

[19]To reduce false positives, matches are restricted to the academic's primary discipline (e.g., physics).

2.3 Socio-Economic Background and the Probability of Becoming an Academic

In the first part of the paper, we investigate the relationship between socio-economic background and the probability of becoming an academic. Many anecdotes suggest that even exceptionally talented individuals from lower socio-economic backgrounds often face challenges in pursuing academic careers. For example, in his *Recollections*, Nobel Prize winner George Beadle stated that: "It was tacitly assumed I would eventually take over the family farm. [...] Father was not keen on the college idea, being convinced that a farmer did not need all that education. But determination won, and I enrolled at the University of Nebraska College of Agriculture, fully intending to return to the farm" (Beadle, 1974).

In the following, we explore whether individuals like George Beadle represent rare exceptions or if talented individuals were able to pursue academic careers regardless of their socio-economic background.

2.3.1 Representation of Academics by Socio-Economic Background

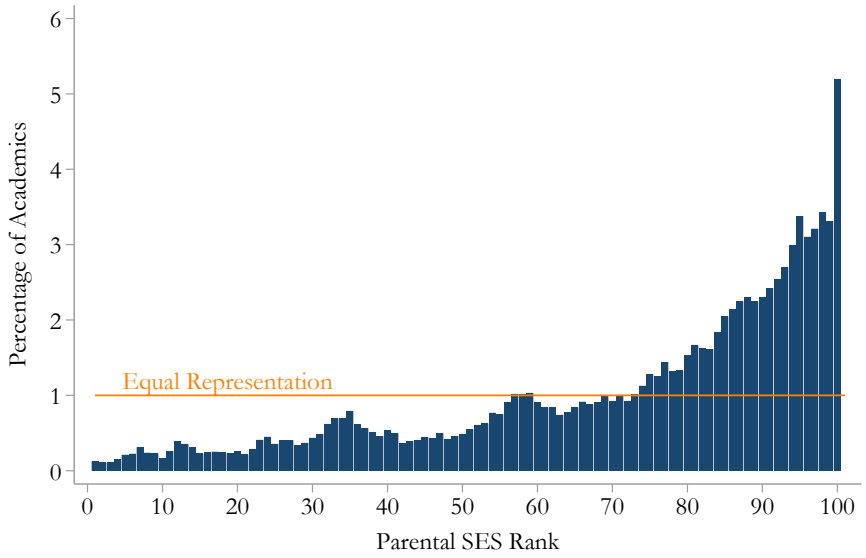
We visualize the share of U.S. academics that come from each percentile of the parental SES rank distribution (Figure 2.3.3). It is important to note that the parental SES rank should be interpreted as an omnibus measure of socio-economic background capturing a combination of different factors such as parental income but also education and other traits of the socio-economic background that are correlated with income. We do not argue that any single factor, such as a lack of parental income, is the sole or even dominant driver of our findings.

An equal distribution based on parental SES ranks would imply that 1% of academics stem from each percentile. We illustrate this benchmark with a horizontal line in Figure 2.3.3. In stark contrast to this equal representation benchmark, we show that people from higher socio-economic backgrounds are markedly overrepresented in academia, with the degree of overrepresentation increasing particularly sharply for higher parental SES ranks (Figure 2.3.3, panel a). Overall, approximately half of all academics come from the top 20% of the parental SES rank distribution. The degree of overrepresentation is particularly large for very high percentiles of the parental SES rank distribution. For example, individuals born to parents in the 95th percentile are more than three times as likely to become academics than one would expect under the equal representation benchmark.

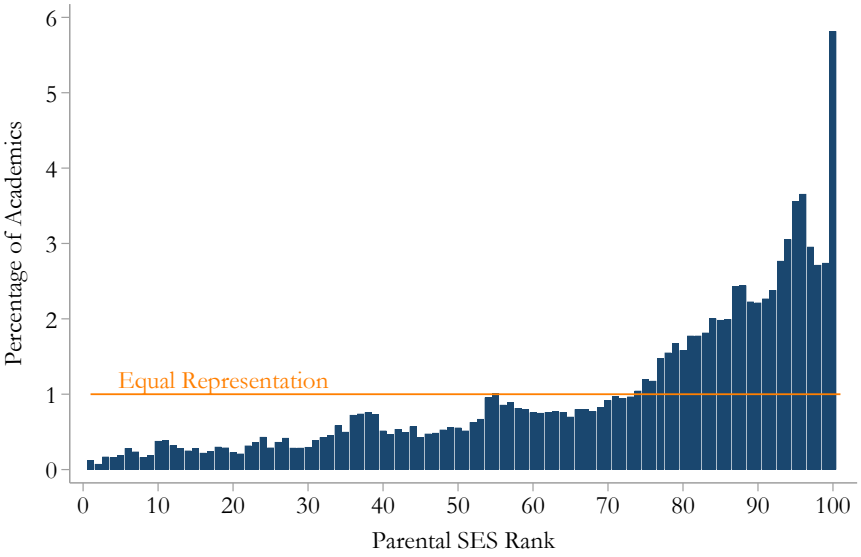
The disparity is even more striking at the highest percentile. Individuals from the 100th percentile of the socio-economic background distribution are more than five times as likely to become academics than one would expect under the equal representation benchmark. Strikingly,

Figure 2.3.3: Representation by Socio-Economic Background

(a) Baseline Parental Income Prediction



(b) Parental Income Prediction Without Regional Variation



Notes: The figure shows the representation of academics based on their socio-economic background for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each bar represents the percentage of all academics whose fathers are from a specific income percentile rank. For example, the right-most bar shows that around 5 percent of academics have fathers who were in the 100th percentile of the predicted income distribution. The horizontal line represents a hypothetical equal representation benchmark. Appendix Figure B.2 shows equivalent figures for the extended sample.

even when compared to individuals from the 99th percentile, those from the 100th percentile have a 1.6 times higher chance of becoming an academic.[20]

The results are similar if we predict parental SES ranks solely based on the father's individual characteristics and his occupation, excluding state of residence fixed effects in the income prediction (Figure 2.3.3, panel b). In additional robustness checks, we report the share of academics by other measures of socio-economic background (Hisclass and Duncan Socioeconomic Index (SEI), Appendix Figures B.3 and B.4) and confirm that academics are disproportionately drawn from high socio-economic backgrounds.

2.3.2 Representation Over Time

The large differences in the probability of becoming an academic translate into a highly skewed socio-economic composition of academia. As a next step, we analyze whether these representation patterns changed over time (Figure 2.3.4). The share of academics from the top quintile of the parental SES rank distribution for the birth cohorts born after 1920 is 52.6%, almost identical to the share of 52.3% in the pre-1870 birth cohorts. Similarly, the share of academics from the bottom quintile of the parental SES rank distribution is around 4-5% and hardly changes over time. This persistence is striking, given the substantial expansion in educational attainment in the United States during this period.[21]

Together, these results suggest that there are significant and persistent barriers that prevent individuals from low socio-economic backgrounds from pursuing careers in academia. Such barriers could take many different forms (e.g., differences in ability, education, income, network ties, or institutional knowledge).

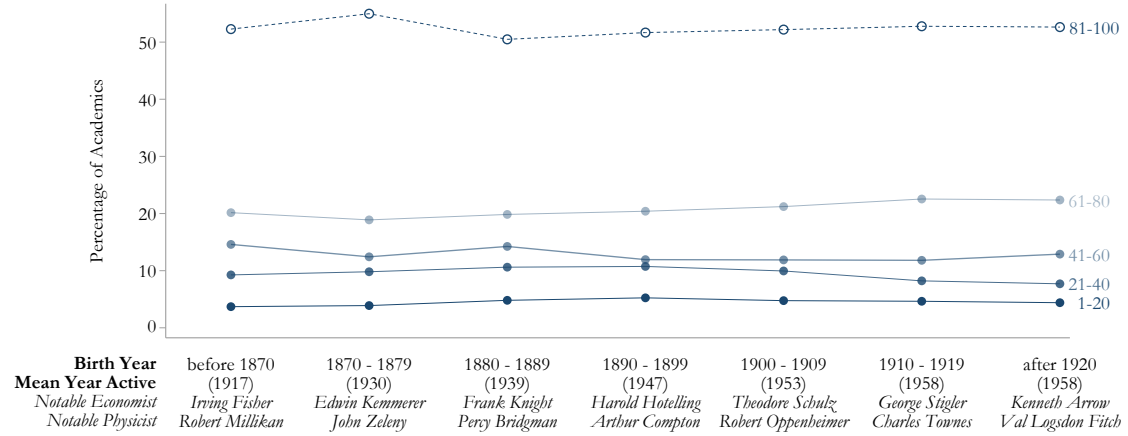
2.3.3 Representation in Academia versus Other Professions

A question arising from the previous findings is whether academia is an outlier compared to other professions. The small share of individuals from low socio-economic backgrounds in academia might simply reflect the fact that entering a profession requires credentials (e.g., a college degree), which might be expensive to obtain. To explore this, we compare the socio-economic backgrounds of academics to those of other professionals – lawyers and judges, physicians and surgeons, and teachers – using comparable data from the census (see Appendix 2.A.2 for details). While lawyers and doctors also disproportionately come from high socio-economic backgrounds, the degree of

[20] This extreme overrepresentation at the 100th percentile may partially reflect that, in certain census years and states, professors themselves are classified in the highest parental income percentile. However, even after excluding individuals whose fathers report "professor" as their occupation in the census, the overall pattern remains similar (Appendix Figure B.1).

[21] For example, U.S. Americans born in 1920, on average, completed three additional years of schooling compared to those born in 1870 (Goldin and Katz, 2009).

Figure 2.3.4: Representation by Socio-Economic Background Over Time



Notes: The figure shows the representation of academics based on their socio-economic background over time for the main sample. Each line represents the percentage of all academics whose fathers are from a specific income quintile. For example, the top line indicates the percentage of academics whose fathers were in the top quintile of the predicted income distribution. Appendix Figure B.5 shows the equivalent figure for the extended sample.

selection in academia is even more pronounced (Figure 2.3.5). For example, 52% of academics come from the top quintile of the parental SES rank distribution, while "only" 50% of laywers and judges, and 44% of medical doctors come from the top quintile of the parental SES rank distribution. At the other end of the spectrum, representation from the bottom quintile of the parental SES rank distribution is especially low in academia: only 5% of academics come from the bottom quintile, while 7% of laywers, and 9% of doctors come from the bottom quintile. Teachers, in contrast, exhibit a much weaker degree of selection based on socio-economic background.

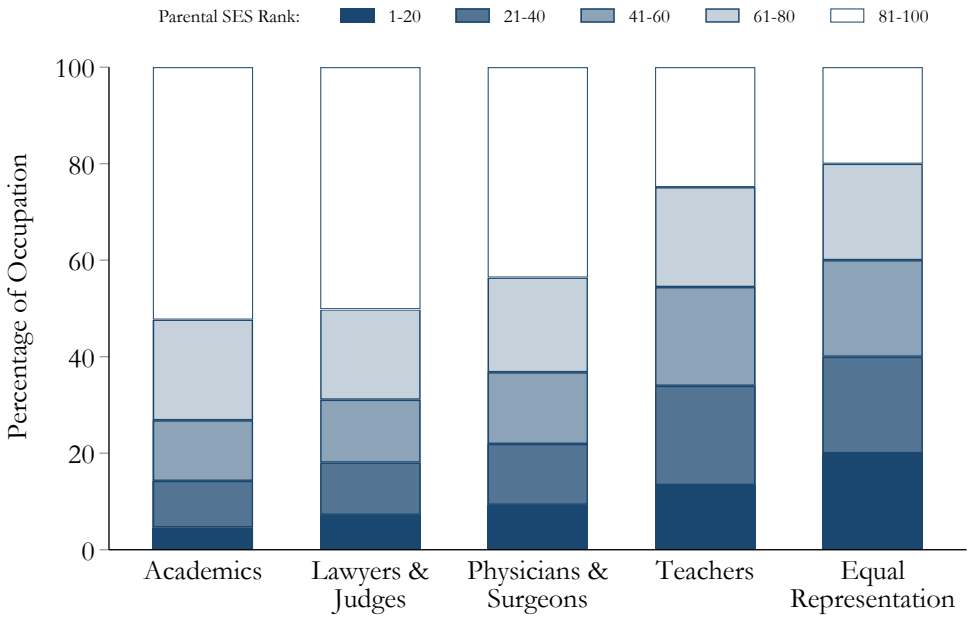
2.3.4 Representation by University

In the next set of results, we investigate whether individuals from lower socio-economic backgrounds are similarly underrepresented in all universities or if certain universities exhibit a higher degree of representation of individuals from these backgrounds. As the faculty rosters contain more than 1,000 U.S. universities, we show examples for a small subset of these universities. We choose examples of universities for which we measure the socio-economic background of academics in each of the five cohorts plus all universities in the Ivy Plus group, as defined by Chetty et al. (2020).[22]

We find striking differences in representation by university (Figure 2.3.6, which is sorted in descending order based on the proportion of faculty with fathers from the top 20%). The most "socio-economically selective" universities are elite private universities such as those in the Ivy

[22]The Ivy Plus group contains the following universities: Brown, Columbia, Cornell, Dartmouth, Harvard, UPenn, Princeton, Yale, Stanford, MIT, Chicago, and Duke.

Figure 2.3.5: Comparison to other Professions



Notes: The figure compares the representation of academics based on their socio-economic background to the representation in other professions for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each color shows the percentage of individuals in an occupation whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of individuals whose father was in the top quintile of the predicted income distribution. The representation in other professions is based on U.S. census samples of lawyers & judges, physicians & surgeons, and teachers that match the sample of academics (see Appendix 2.A.2 for details). Appendix Figure B.6 shows the equivalent figure for the extended sample.

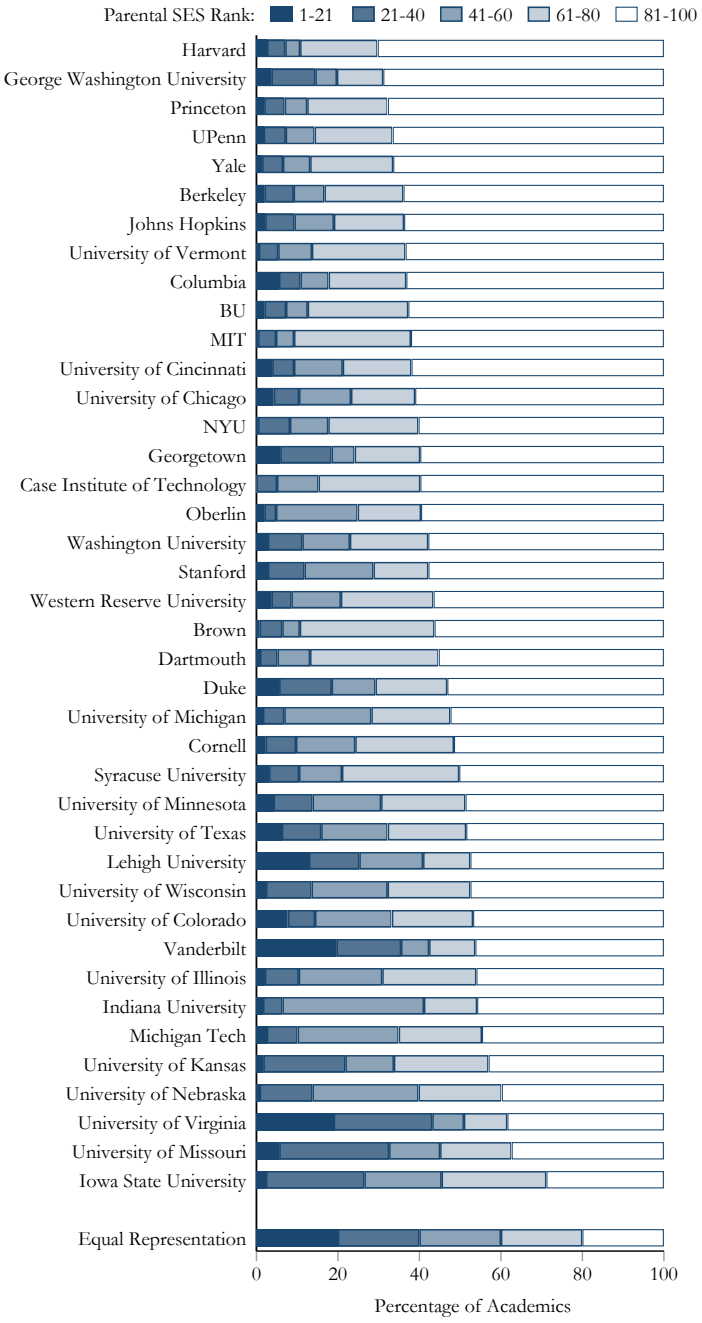
League – Harvard, Princeton, UPenn, and Yale. In contrast, universities with lower levels of "social selectivity" within this subset are predominantly public institutions, such as the University of Nebraska, the University of Missouri, and Iowa State University. These differences highlight significant variation in socio-economic representation across universities.

To more systematically investigate which university characteristics are correlated with socioeconomic selectivity, we estimate the following regression on the full sample of universities:

$$\begin{aligned}\text{Faculty Top SES Share}_u = \beta_0 + \beta_1 \text{Ivy Plus}_u + \beta_2 \text{Elite Private}_u + \beta_3 \text{Elite Public}_u \\ + \beta_4 \text{Discipline Shares}_u + \text{State FE} + \epsilon_i\end{aligned} \tag{2.3.2}$$

The dependent variable Faculty Top SES Share$_u$ measures the share of academics of university u who come from the top 20, top 10, top 5, or top 1 % of the parental SES rank distribution.

Figure 2.3.6: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top quintile of the predicted income distribution. Appendix Figure B.7 shows the equivalent figure for the extended sample.

Ivy Plus$_u$ is an indicator that equals one if university u is an Ivy Plus university as defined by Chetty et al. (2020). Elite Private$_u$ is an indicator that equals one if university u is an elite private institution which is not in the Ivy Plus category (e.g., New York University) and Elite Public$_u$ is an indicator that equals one if the university is an elite public institution (e.g., Berkeley).[23]

The regression results indicate that Ivy Plus universities recruit faculty from significantly higher socio-economic backgrounds compared to other elite private institutions. These findings hold for the share of faculty from the top 20, top 10, top 5, and even top 1 %. While the average university in our sample recruits 3.4 % of their academics from the top 1 %, the share is about 5.2 percentage points higher in Ivy Plus universities (Table 2.3.3, column 12). In contrast, public elite institutions recruit their faculty from lower socio-economic backgrounds than Ivy Plus universities (Table 2.3.3).

The selectivity of universities may, in part, reflect their discipline composition. For example, Harvard does not have an agriculture department, which could influence the selectivity of its faculty. As demonstrated in the next section, representation varies substantially across disciplines. To address these differences, we add controls for the share of academics in each discipline. The results remain very similar (columns 2, 5, 8, and 11). The differences across university types are similar even though somewhat smaller if we control for state fixed effects (columns 3, 6, 9, and 12). This suggests that the observed patterns are not solely driven by geographical factors.

2.3.5 Representation by Discipline

While individuals from higher socio-economic backgrounds are overrepresented in all disciplines, there are large differences across disciplines (Figure 2.3.7). Agriculture, veterinary medicine, pedagogy, sociology, and pharmaceutics are the disciplines with the highest representation of individuals from lower socio-economic backgrounds. In contrast, the humanities, archaeology, architecture, cultural studies, medicine, anthropology, and law have the lowest representation.[24] Contrary to the common perception of economists, economics is more representative than the median discipline.

Figure 2.3.7 suggests that disciplines that require more sophisticated language skills have less representation from individuals of lower socio-economic backgrounds. In comparison, disciplines that require more mathematics skills exhibit higher representation. To investigate this hypothesis, we correlate discipline-level representation with the language versus mathematics skills requirement in each discipline. We proxy the language versus mathematics requirement with the ratio of quantitative to verbal Graduate Record Examination (GRE) scores for students intending to

[23] Elite Private includes all private universities in Chetty et al. (2020)'s "elite universities". Elite Public includes all public universities in Chetty et al. (2020)'s "elite universities" as well as all universities in their "Highly-Selective Public" category.

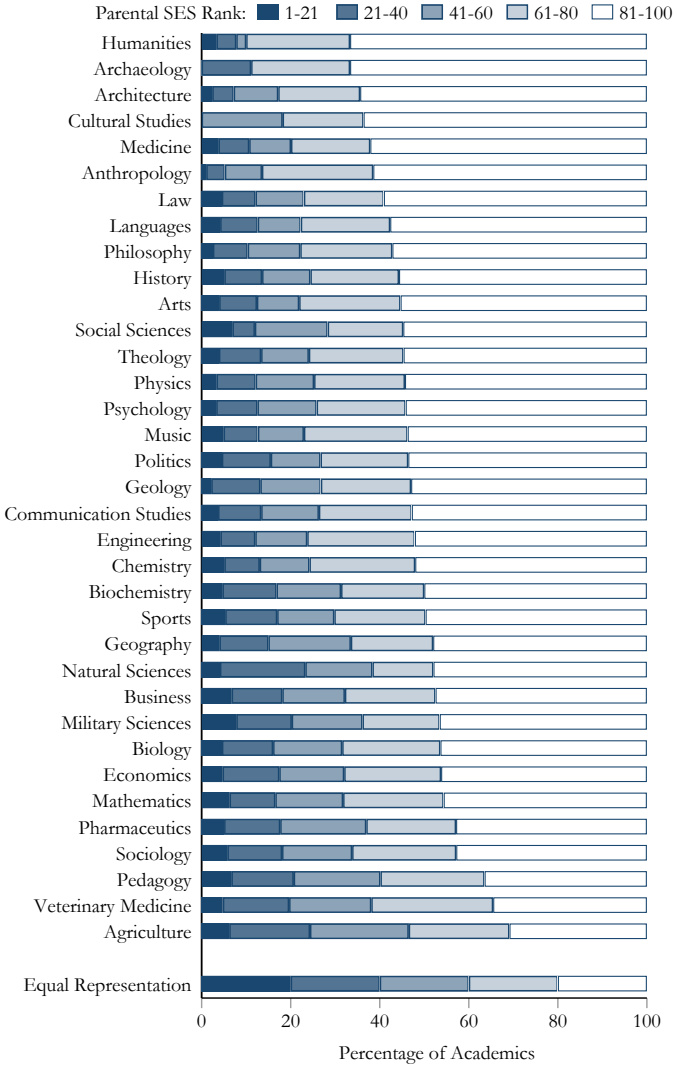
[24] Academics who list humanities, social sciences, and natural science as their discipline in the faculty rosters are less specialized and often teach at liberal arts colleges.

Table 2.3.3: Correlates of University SES-Selectivity

Dependent Variable:	*Faculty Top SES Share*											
	20%			*10%*			*5%*			*1%*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: 1900 – 1956												
Ivy Plus	0.103**	0.096***	0.037	0.135***	0.115***	0.068***	0.115***	0.106***	0.054***	0.058***	0.059***	0.052***
	(0.045)	(0.025)	(0.028)	(0.031)	(0.020)	(0.020)	(0.023)	(0.019)	(0.017)	(0.008)	(0.020)	(0.019)
Private Elite	0.100***	0.059*	0.032	0.093***	0.054**	0.032	0.078***	0.044**	0.030	0.025***	0.012	0.010
	(0.031)	(0.031)	(0.035)	(0.023)	(0.023)	(0.026)	(0.016)	(0.016)	(0.018)	(0.007)	(0.009)	(0.009)
Public Elite	0.083***	0.092**	0.076*	0.028	0.027	0.030	0.019	0.015	0.018	0.011*	0.007	0.007
	(0.029)	(0.037)	(0.041)	(0.019)	(0.023)	(0.031)	(0.015)	(0.018)	(0.022)	(0.006)	(0.007)	(0.011)
R^2	0.02	0.10	0.19	0.02	0.10	0.18	0.03	0.13	0.22	0.03	0.10	0.16
Observations	755	755	755	755	755	755	755	755	755	755	755	755
Dependent Variable Mean	0.481	0.481	0.481	0.270	0.270	0.270	0.138	0.138	0.138	0.034	0.034	0.034
Panel B: 1900 – 1969												
Ivy Plus	0.146***	0.111***	0.059**	0.153***	0.124***	0.077***	0.112***	0.098***	0.054***	0.048***	0.044***	0.031**
	(0.040)	(0.031)	(0.030)	(0.031)	(0.026)	(0.023)	(0.029)	(0.018)	(0.014)	(0.005)	(0.014)	(0.014)
Private Elite	0.114***	0.056*	0.039	0.097***	0.049*	0.030	0.071***	0.039**	0.029	0.020***	0.008	0.009
	(0.031)	(0.031)	(0.037)	(0.025)	(0.025)	(0.029)	(0.014)	(0.017)	(0.018)	(0.006)	(0.010)	(0.010)
Public Elite	0.057	0.002	0.014	0.043	0.000	0.024	0.033	0.003	0.017	0.003	-0.006	-0.012
	(0.045)	(0.033)	(0.031)	(0.035)	(0.032)	(0.023)	(0.026)	(0.023)	(0.017)	(0.007)	(0.007)	(0.008)
R^2	0.01	0.08	0.16	0.02	0.10	0.18	0.02	0.08	0.17	0.01	0.05	0.10
Observations	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026
Dependent Variable Mean	0.449	0.449	0.449	0.249	0.249	0.249	0.130	0.130	0.130	0.035	0.035	0.035
Discipline Controls		Yes	Yes		Yes	Yes		Yes	Yes		Yes	Yes
State FEs			Yes			Yes			Yes			Yes

Notes: The table reports the estimates of Equation (2.3.2). The dependent variable measures the share of faculty in university u who come from the top 20 (columns 1-3), top 10 (columns 4-6), top 5 (columns 7-9), or top 1 percent (columns 10-12) of the parental SES rank distribution, respectively. Ivy Plus_u is an indicator that equals one if university u is an Ivy Plus university as defined by Chetty et al. (2020). Elite Private_u is an indicator that equals one if university u is an elite private institution which is not in the Ivy Plus category (e.g., New York University) and Elite Public_u is an indicator that equals one if university u is an elite public institution (e.g., Berkeley). Standard errors are clustered at the state-level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Figure 2.3.7: Representation by Discipline



Notes: The figure shows the representation of academics based on their socio-economic background by discipline for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top quintile of predicted income. Appendix Figure B.8 shows the equivalent figure for the extended sample.

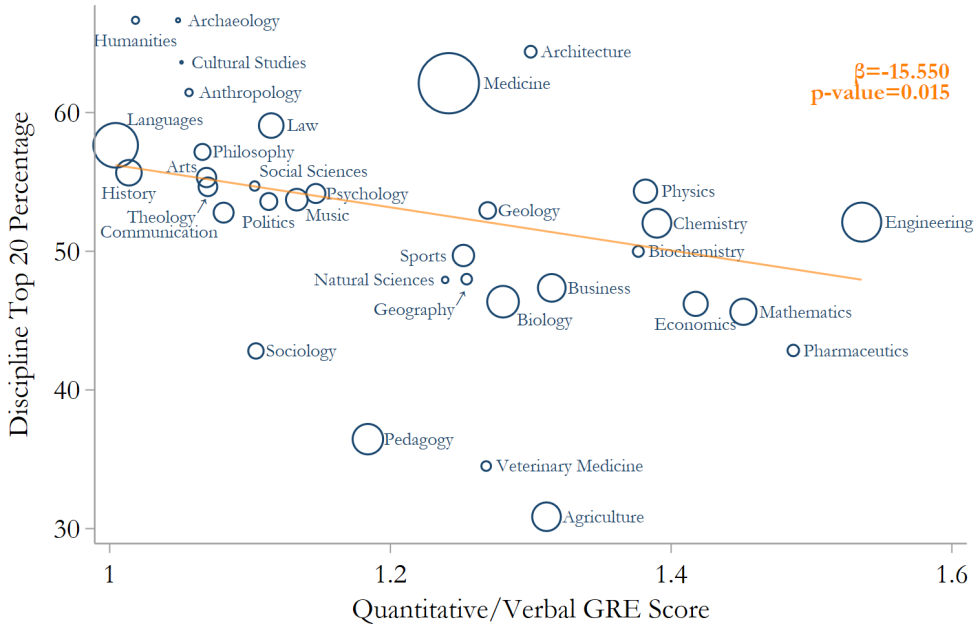
pursue graduate studies in each discipline.[25] The findings suggest that representation from lower socio-economic backgrounds is indeed higher in disciplines that require more quantitative relative

[25]The Educational Testing Service (ETS), which administers the GRE, publishes three-year average test scores of seniors and nonenrolled college graduates in three categories (verbal reasoning, quantitative reasoning and analytical writing) for 290 intended graduate majors in their *GRE Guide to the Use of Scores*. We aggregate these majors into the corresponding disciplines from the faculty rosters and calculate the average quantitative versus verbal GRE score in each discipline. The data used in this analysis is based on the 2005–2008 cohorts of test-takers, obtained from the oldest available edition of the guide available via the Internet Archive Wayback Machine (ETS, 2009).

to verbal skills (Figure 2.3.8). The estimates imply that an increase in relative quantitative versus verbal skills by 0.5 (approximately the difference between history and mathematics) is associated with a 7.8 percentage point decrease in the share of academics from the top quintile of the parental SES rank distribution.

However, Figure 2.3.7 also highlights striking differences in representation even when comparing disciplines that arguably require similar skills. For instance, there are large differences in the socio-economic composition of medicine relative to veterinary medicine and of sociology relative to law. This suggests that factors beyond skill requirements also impact representation across disciplines.

Figure 2.3.8: Discipline Mathematics vs. Language Requirements and Representation



Notes: The figure shows the share of academics from the top quintile of the distribution of socio-economic background by academic discipline in relation to the importance of quantitative relative to verbal skills in the discipline for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. We proxy the importance of mathematics relative to language skills with the ratio of the average GRE quantitative score to the average GRE verbal reasoning score of test takers intending to pursue a graduate degree in the respective discipline. GRE score data come from ETS (2009), Extended Table 4. The size of the circles indicates the number of academics in the respective discipline in our data. We also report the coefficient and p-value from a discipline-size weighted regression of this relationship. Appendix Figure B.9 shows the equivalent figure for the extended sample.

2.4 Socio-Economic Background and Discipline Choice

In the second part of the analysis, we examine whether fathers' occupation affects academics' choice of discipline. This enables us to study a different facet of socio-economic background that goes beyond fathers' positions in the SES rank distribution.

2.4.1 Measuring Discipline-Level Overrepresentation by Father's Occupation

For this analysis, we construct an overrepresentation index that measures whether individuals with fathers in certain occupations are overrepresented in specific academic disciplines:

$$\text{Overrepresentation}_{do} = \frac{P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o)}{P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)}, \tag{2.4.3}$$

where $P(\text{Discipline}_i = d)$ is the probability of academic i working in discipline d, $P(\text{Father's Occupation}_i = o)$ is the probability of academic i having a father with occupation o, and $P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o)$ is the joint probability.[26]

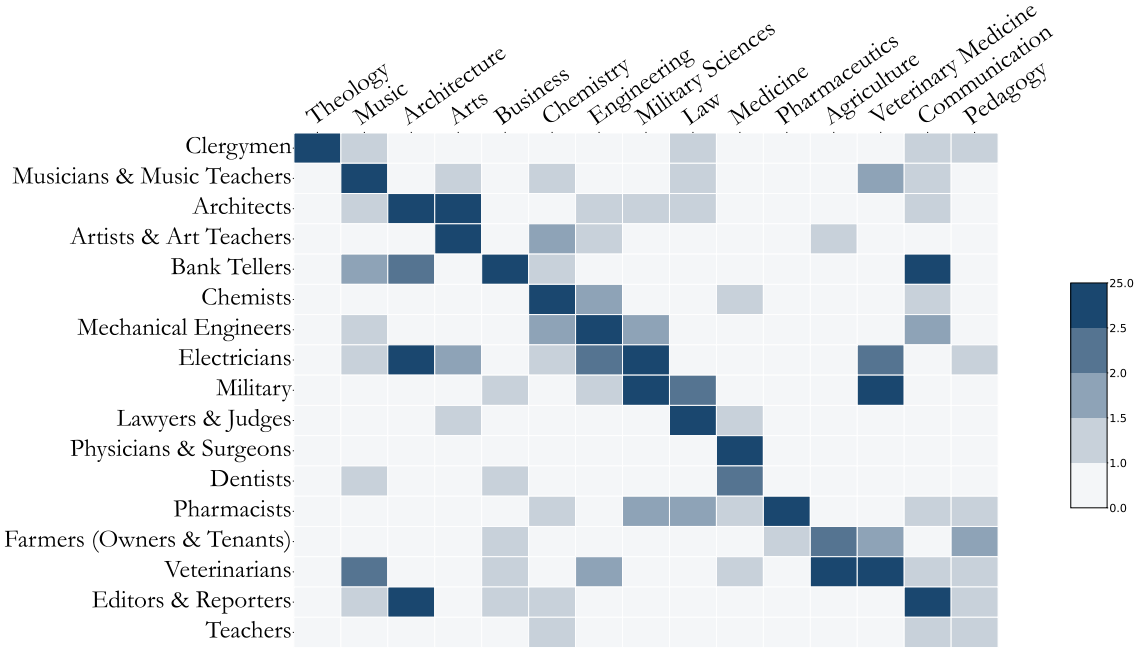
The measure isolates the relationship between a father's occupation and an academic discipline by accounting for baseline differences in the probabilities of choosing specific disciplines and having fathers in certain occupations. If there was no systematic relationship between father's occupation and the choice of discipline (i.e., the probabilities are independent), $\text{Overrepresentation}_{od} = 1$, since $P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o) = P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)$. If a certain father's occupation is overrepresented in a specific discipline, the measure is greater than one. Inversely, in case of underrepresentation, the measure is smaller than one.

For example, we can calculate the overrepresentation of farmers' children among professors of agricultural science. The numerator measures the probability that an academic whose father was a farmer works as a professor of agricultural science (in our data this probability is 0.024). The denominator is the product of two probabilities: the probability of being a professor of agriculture among all academics (in our data: 0.043), and the probability that any academic's father was a farmer (in our data: 0.232). Thus the overrepresentation index for professors of agriculture who are farmer's children is $0.024/(0.043 \cdot 0.232) = 2.4$. In other words, 56% ($0.024/0.043 \times 100$) of all agricultural scientists are the children of farmers, while only 23% of all academics are children of farmers, making agricultural scientists 2.4 times more likely to be the child of a farmer, compared to academics overall. Thus, the measure quantifies the extent to which children of farmers are disproportionately represented in agricultural sciences.

[26] The measure is related to pointwise mutual information, a common measure in information theory.

We calculate this measure for all pairs of father's occupations (130 different occupations in the data) and academic disciplines (34 disciplines), i.e., we calculate $130 \times 34 = 4,420$ overrepresentation indices.[27] We visualize examples of such pairs in Figure 2.4.9. The figure plots the father's occupation on the vertical axis and the academic discipline on the horizontal axis. The blue shading indicates quartiles of the overrepresentation index, with darker blues indicating stronger overrepresentation.

Figure 2.4.9: Father's Occupation and Discipline Choice



Notes: The figure shows the relationship between father's occupation (rows) and their children's academic discipline choice (columns) for selected father's occupation - discipline pairs for the main sample. Darker shades indicate higher levels of overrepresentation, as measured by equation (2.4.3). Appendix Figure C.1 shows the equivalent figure for the extended sample.

The figure suggests a strong connection between the father's occupation and their children's choice of discipline. For example, children of architects are disproportionately represented in architecture and arts, while children of artists and art teachers gravitate toward arts-related disciplines. Children of lawyers, medical doctors, or pharmacists predominantly pursue law, medicine, and pharmaceutics, respectively. Children of editors and reporters are overrepresented in communication studies, which encompasses journalism as a sub-discipline. Interestingly, this pattern extends to children of fathers in non-professional occupations. For example, children of bank tellers are overrepresented in business disciplines. Meanwhile, children of teachers, who often teach various school subjects, exhibit a more evenly distributed representation across academic fields.

[27] As the overrepresentation index is sensitive to outliers in small disciplines and occupations, we restrict the sample to disciplines for which we can observe the occupation of the father for at least 15 academics and to fathers' occupations in which at least 15 children become academics in any discipline.

2.4.2 Predicting Semantically Close Academic Disciplines

Figure 2.4.9 presents a selected subset of father's occupation-discipline pairs, that we hand-picked from the data chosen to illustrate notable patterns in the data. To systematically investigate the relationship between a father's occupation and their child's academic discipline choice, we construct an external measure of similarity between each father's occupation and each academic discipline. Specifically, we use text embeddings to measure the *semantic* similarity between the text string of the father's occupation and the text string of the discipline. This method provides a systematic way to explore the relationship between father's occupation and the discipline for all father's occupation-discipline pairs.

Embeddings transform a text into a fixed-length vector representation that capture both syntactic and semantic relationships present in the training data. The resulting vectors can then be used for text similarity calculations, as similar sentences are located close to each other in the vector space. Intuitively, if the word "farmer" is used in similar contexts to the word "agriculture", the model will identify these words as being semantically similar. Embedding models are trained by applying advanced machine learning techniques, such as deep learning transformer models, to vast corpora of text such that the model learns intricate relationships between words. We use the "all-MiniLM-L6-v2" model, which has been trained on data from scientific papers, Wikipedia, Reddit, and many other sources.[28] The model represents each father's occupation string as well as each discipline string as a vector of length $n = 384$. As is standard in natural language processing, we then measure the similarity of the text string of the father's occupation and the text string of the discipline using the cosine similarity of the two vector representations:

$$\text{Cosine Similarity(x,y)} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{2.4.4}$$

where x represents the vector of father's occupation and y represents the vector of the discipline, derived from the sentence embedding model.

Using this measure of semantic similarity, we predict the closest discipline in semantic space for each father's occupation. Importantly, this measure is derived solely from the textual representation of occupation and discipline *strings* and does not incorporate any information about the actual academic discipline choices of professors. For example, as expected the closest discipline in semantic space for the occupation "architect" is "architecture" (cosine similarity 0.77). Similarly, the closest discipline in semantic space for the occupation "buyers and shippers, farm products" is "agriculture" (cosine similarity 0.53).[29]

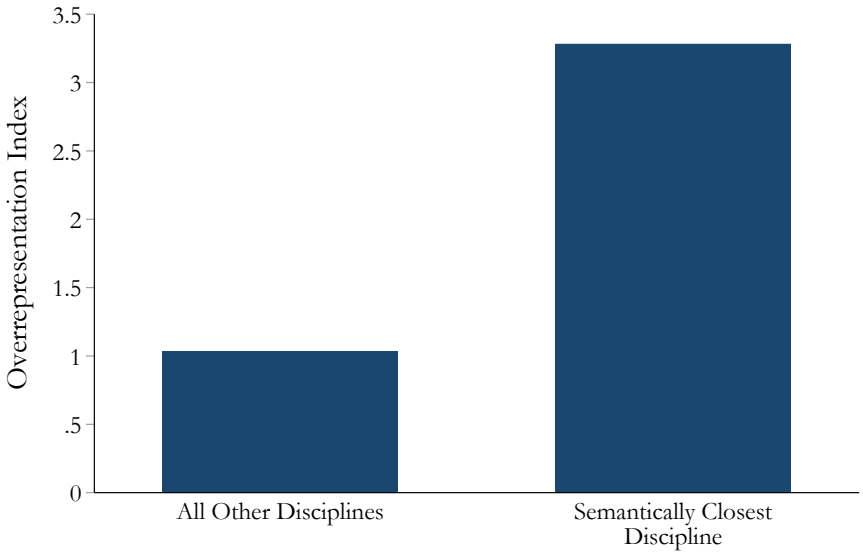
[28]The "all-MiniLM-L6-v" model is one of the most commonly used sentence embedding models. For example, it was the third most downloaded model on huggingface.com as of July 2024. The findings do not depend on the choice of a specific model.

[29]To ensure that we predict close disciplines that are genuinely close in semantic space, we classify an occupation-

2.4.3 Overrepresentation in Semantically Close Disciplines

After identifying the semantically closest academic discipline for each occupation, we compute the average overrepresentation index (equation 2.4.3) for the discipline-occupation pair that is closest in semantic space. Additionally, we calculate the corresponding average for all other discipline-occupation pairs. This enables us to measure whether academics are *systematically* overrepresented in disciplines that are "close" to their father's occupation.

Figure 2.4.10: Overrepresentation in Semantically Closest Discipline



Notes: The figure shows overrepresentation as measured by equation (2.4.3) in the father's occupation-discipline pair that is semantically closest, e.g., "farmer" and "agriculture" and all other father's occupation-discipline pairs for the main sample. For more details, see appendix 2.4.2 and appendix 2.4.3. Appendix Figure C.2 shows the equivalent figure for the extended sample.

The average overrepresentation index is 3.28 in the semantically closest discipline (Figure 2.4.10). In contrast, the overrepresentation index is 1.03 for all other disciplines, indicating that for disciplines that are not semantically close to fathers' occupations, academics are represented as good as random.

Overall, these results provide further evidence that socio-economic background not only affects the likelihood of entering academia but also the choice of discipline. Potential explanations for this phenomenon include increased interest stemming from the transmission of family values, early exposure to a particular field, or differential access to resources and opportunities, such as privileged

discipline pair as semantically close if their cosine similarity is at least two standard deviations above the mean of all cosine similarities. For instance, while the discipline most similar to the occupation "private household worker" is law, the similarity falls below the mean cosine similarity threshold. As a result, children of "private household workers" have no semantically closest discipline and are excluded from our main analysis. Importantly, our findings remain robust when we redefine semantically close disciplines using a threshold of one standard deviation above the mean or when we eliminate the minimum cosine similarity requirement altogether (see Appendix Figure C.3).

information on how to succeed in a given discipline.

Combined with the findings in the previous part of the paper, these results suggest that the unequal selection of academics based on socio-economic background could have repercussions for the composition of academic disciplines. Overrepresentation of individuals from certain parental occupations in academia could skew the composition of academic disciplines, leading to imbalances in the supply of talent. This misalignment may advantage some disciplines over others, not due to societal demand for knowledge, but rather due to the unequal distribution of opportunities.

2.5 Socio-Economic Background, Scientific Publications, and Novel Scientific Concepts

In the third part of the analysis, we investigate whether and how socio-economic background influences productivity after entering academia. In particular, we study whether scientific productivity and novelty differ by socio-economic background.

2.5.1 Scientific Publications

We first explore differences in the number of publications by socio-economic background. As described in the data section, this analysis focuses on six scientific disciplines: medicine, biology, biochemistry, chemistry, physics, and mathematics, which are well-represented in academic publication databases. We estimate the following regression:

$$Publications_i = \theta \cdot \text{Parental SES Rank}_i + \mathbf{X_i'}\beta + \epsilon_i \tag{2.5.5}$$

where $Publications_i$ captures different measures of scientific publications that scientist i published in a $\pm$ 5-year interval centered on the year that the scientist entered the faculty rosters, i.e., for scientists entering the faculty rosters in 1956, we measure publications from 1951 to 1961. We estimate results for publication counts and standardized publications. We standardize publication counts to have a mean of zero and a standard deviation of one by discipline and cohort.[30] Standardized publications ease interpretation and account for differences in publications across disciplines and over time. Parental SES Rank$_i$ ranges from 0 to 100, capturing the percentile of the income rank of scientist i's father. The coefficient θ captures the relationship between socio-economic background and scientific output. We also include a set of controls, $\mathbf{X_i}$, to account for differences in scientific output by age, gender, cohort, discipline, or state. Since the parental SES rank is based on father's

[30]To capture the whole distribution of publications for the standardization, we use all publications linked to U.S. scientists in the faculty rosters and not only the publications of U.S. scientists which we can link to a childhood census.

occupation, childhood state, and birth year of the scientist, we cluster standard errors at the level of father's occupation, childhood state, and birth year to account for potential correlations of regression residuals.

Number of Publications

We find no systematic relationship between the socio-economic background and the *average* number of publications, regardless of the set of fixed effects that we include as regression controls (Table 2.5.4, columns 1-3). This result holds in the main sample (Panel A) and in the extended sample (Panel B). As described before, to account for differences in publication practices across disciplines and over time, we also estimate models using standardized publications. These results further confirm that there is no systematic relationship between the socio-economic background of scientists and the average number of publications (Table 2.5.4, columns 4-6).

Table 2.5.4: Socio-Economic Background and Publications

Dependent Variable:	*Publications*			*Standardized Publications*			*No Publications*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: 1900 – 1956									
Parental SES Rank	0.00783*	0.00441	-0.00299	0.00040	0.00012	0.00006	-0.00113***	-0.00094***	-0.00052***
	(0.00425)	(0.00424)	(0.00423)	(0.00041)	(0.00041)	(0.00042)	(0.00019)	(0.00019)	(0.00018)
R^2	0.04	0.06	0.09	0.04	0.06	0.06	0.05	0.07	0.12
Observations	12,767	12,767	12,767	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	4.666	4.666	4.666	0.011	0.011	0.011	0.418	0.418	0.418
Panel B: 1900 – 1969									
Parental SES Rank	0.00419	0.00158	-0.00628	0.00016	-0.00005	-0.00014	-0.00102***	-0.00085***	-0.00043**
	(0.00408)	(0.00408)	(0.00407)	(0.00036)	(0.00036)	(0.00036)	(0.00017)	(0.00017)	(0.00017)
R^2	0.03	0.05	0.08	0.03	0.06	0.06	0.04	0.06	0.12
Observations	15,521	15,521	15,521	15,521	15,521	15,521	15,521	15,521	15,521
Dependent Variable Mean	4.912	4.912	4.912	-0.013	-0.013	-0.013	0.421	0.421	0.421
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes			Yes

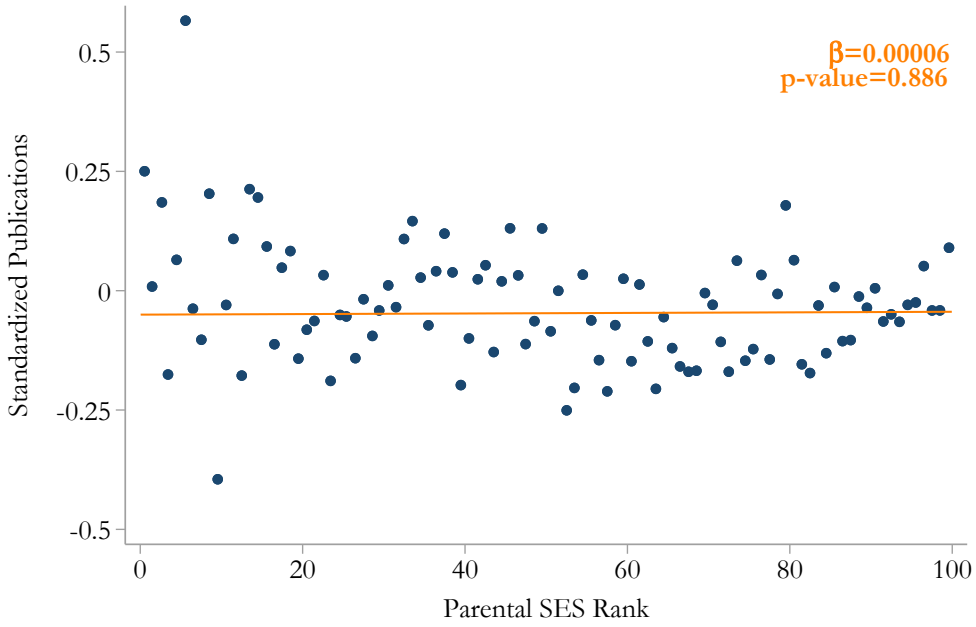
Notes: The table reports the estimates of equation (2.5.5). The dependent variable measures publications in a ± 5-year window around the cohort when scientist i enters the faculty rosters. We standardize publications to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist is female. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the scientist. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

We also visualize the relationship between parental income ranks (x-axis) and standardized publications (y-axis) in a binned scatterplot (Figure 2.5.11). The figure provides additional evidence that there is no systematic relationship between the *average* number of publications and the parental income rank.

Probability of Zero Publications

A considerable share of scientists never publish in journals indexed by the Web of Science, which predominantly includes high-quality journals (Hager et al., 2024). To examine the likelihood of

Figure 2.5.11: Socio-Economic Background and Average Number of Publications



Notes: The figure shows a binned scatterplot of the relationship between scientists' socio-economic background and publications. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Publications are standardized within cohort and discipline. We show 100 quantiles and use the covariate adjustment (equivalent to column (4) in Table 2.5.4) as proposed in Cattaneo et al. (2024).

never publishing in a Web of Science-indexed journal, we estimate variants of equation (2.5.5) with an alternative dependent variable that equals one if scientist i does not publish any papers in the ± 5 year window surrounding their entry into the faculty rosters, and zero otherwise. We find that individuals from higher socio-economic backgrounds are significantly less likely to never publish (Table 2.5.4, column 7). For example, the probability of not publishing at all is approximately 4 percentage points (or around 10 percent) lower for scientists whose fathers were at the 75th percentile of the income distribution, compared to those with fathers at the 25th percentile. While the magnitude of this effect is halved when including the full set of fixed effects, it remains highly significant. The result is also robust in the extended sample (Table 2.5.4, columns 8-9 and Panel B).

The Distribution of Publications

The preceding results suggest that while scientists from lower socio-economic backgrounds, on average, produce a comparable total number of publications, they are significantly more likely to have no publications at all. This suggests that scientists from lower socio-economic backgrounds must publish relatively more in higher percentiles of the publication distribution. To test this

hypothesis, we estimate equation (2.5.5) with alternative dependent variables:

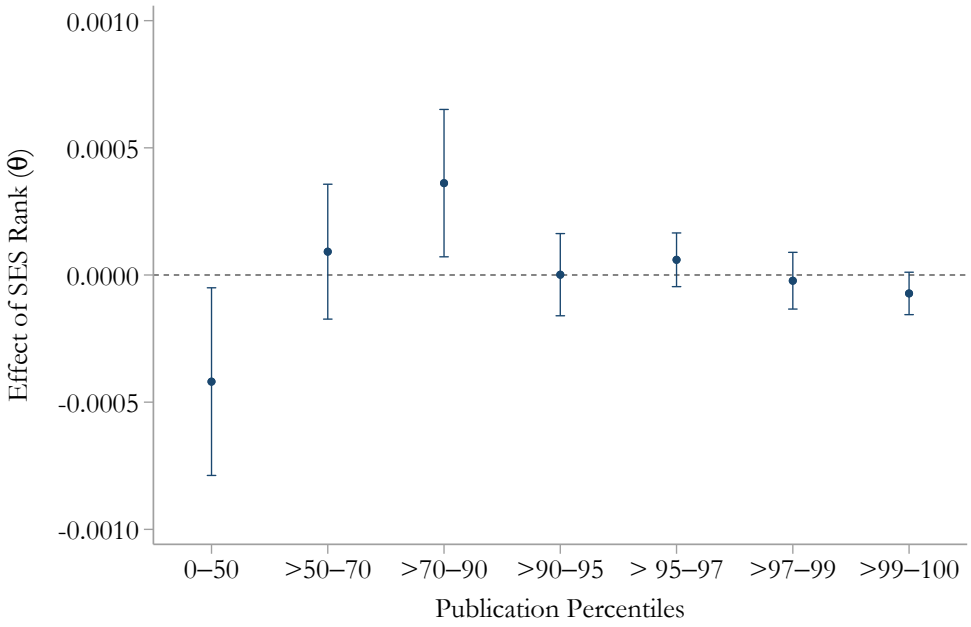
$$\mathbb{1}(\text{Publication Percentile Range}_i = q) = \theta \cdot \text{Parental SES Rank}_i + \mathbf{X_i'}\beta + \epsilon_i \tag{2.5.6}$$

where the dependent variable $\mathbb{1}(\text{Publication Percentile Range}_i = q)$ is an indicator that equals one if scientist i's publication record falls within a specified percentile range q. Since scientific productivity is well-known to be highly skewed (see e.g., Lotka 1926), we define the following percentile ranges of the publication distribution: $0 - 50$th (which coincides with not publishing at all for many disciplines and cohorts), $> 50 - 70$th, $> 70 - 90$th, $> 90 - 95$th, $> 95 - 97$th, $> 97 - 99$th, and > 99th percentile of the publication distribution. To account for variations in publication patterns across disciplines (e.g., chemists and medical researchers publish more than mathematicians) and cohorts (e.g., later cohorts tend to publish more), these percentiles are calculated at the discipline-cohort level. Appendix Table D.1 shows the number of publications required to achieve each percentile across disciplines and cohorts.

The regression results are reported in Appendix Table D.2, and the estimated coefficients are visualized in Figure 2.5.12. The first coefficient from the left indicates that scientists from higher parental income ranks are less likely to have a publication count in the bottom 50% of the publication distribution. In contrast, the second coefficient ($> 50 - 70$) indicates that scientists from higher parental income ranks are as likely as scientists from lower parental income ranks to have a publication count between the 50th and the 70th percentile of the publication distribution. The third coefficient ($> 70 - 90$) indicates that scientists from higher parental income ranks are more likely to have a publication count between the 70th and the 90th percentile of the publication distribution than scientists from lower parental income ranks. For the next percentile ranges, the coefficients are not significantly different from zero. In contrast, the last coefficient ($> 99 - 100$), indicates that scientists from higher parental income ranks are less likely to have a publication count in the top 1% of the publication distribution (p-value=0.089). This suggests that individuals from lower socio-economic backgrounds are disproportionately more likely to have publication records in the top 1%. Specifically, the probability of having a publication record in the top 1% is approximately 0.35 percentage points (or around 44 percent) lower for scientists whose fathers were at the 75th percentile of the income distribution compared to scientists with fathers at the 25th percentile. This large effect, in percentage terms, is particularly relevant as a long-standing literature in the sociology of science has highlighted that the most productive scientists have a disproportionate impact on the advancement of science (e.g., Lotka 1926, Merton 1957).

Overall, the findings on the distribution of publications suggest that scientists from lower socio-economic backgrounds may represent relatively "riskier" hires. They are more likely to have

Figure 2.5.12: Socio-Economic Background and the Distribution of Publications



Notes: The figure plots the estimated coefficients for θ for seven regressions of Equation 2.5.6. In each of the seven regressions, the dependent variable is an indicator of whether a scientist's number of publications falls within the relevant percentiles of the publication distribution, measured at the cohort and discipline-level. We report coefficients from regressions using the covariates and fixed effects equivalent to column (3) in Table 2.5.4. The corresponding regression results are reported in Appendix Table D.2.

no publications at all but are also disproportionately represented in the top 1% of the publication distribution.

2.5.2 Novel Scientific Concepts

In the next subsection, we explore whether and how the content of publications differs by socio-economic background and explore additional evidence whether scientists from lower socio-economic backgrounds may pursue riskier research agendas.

To explore these hypotheses, we adopt the methodology developed by Iaria et al. (2018) to measure the number of novel words introduced by a scientist to the scientific community. The measure proxies for the introduction of new scientific concepts that required novel scientific terms. Specifically, we define novel words as words that were first used in the title of a paper and had not been used in the title of any prior paper included in the entire Web of Science database (not just the papers published by the scientists in our sample).

As the coverage of the Web of Science begins in 1900, we compute the novel words measure for paper titles published from 1910 onwards. This approach allows for a 10-year window to identify

words appearing in scientific papers before designating a term as novel. Consequently, we cannot measure the introduction of novel words for scientists who enter the faculty rosters in 1900. To ensure that we do not consider words that were already in use in other domains, we exclude frequently used words, as well as numbers, from the data.[31]

One example of a novel scientific term is "microbeam," which was used and developed by Raymond E. Zirkle to study the effects of ionizing radiation on living cells. Zirkle, who is widely regarded as the pioneer in the field of radiation biology, grew up on a farm in northern Oklahoma. "As a young boy, his only source of education were one-room country schoolhouses in Oklahoma and southern Missouri. He gained exposure to the outside world and science through reading books." (Atomic Heritage Foundation, 2022). During WW2, Zirkle became a principal investigator in the Manhattan Project biological research program, where he worked on assessing the risk of radiation. From 1944 onwards, he worked at the University of Chicago, where he served as director of the Institute of Radiobiology and Biophysics. In 1952, he also became the first president of the Radiation Research Society.

To examine how socio-economic background is associated with the introduction of novel scientific terms, we estimate the following regression:

$$\text{Novel Words}_i = \omega \cdot \text{Parental SES Rank}_i + \mathbf{X_i'}\beta + \epsilon_i \tag{2.5.7}$$

where Novel Words$_i$ measures the number of papers with at least one novel word that scientist i published in the $\pm$ 5-year interval around entering the faculty rosters. For example, for scientists entering the faculty rosters in 1956, we measure the number of papers published between 1951 and 1961 that introduced at least one novel word. To facilitate interpretation, and to account for differences in the number of novel words introduced in different disciplines and over time, we standardize novel word counts by discipline and cohort to have a mean of zero and a standard deviation of one. As before, Parental SES Rank$_i$ ranges from 0 to 100 and measures the percentile of the income rank of scientist i's father. $\mathbf{X_i}$ are controls that account for differences in introducing novel words by age, cohort, and discipline.

The baseline specification controls for age, gender, childhood state fixed effects, and cohort fixed effects. We find that scientists from higher socio-economic backgrounds introduce fewer novel words

[31]We exclude the 20,000 most frequently used words in English-language books contained in the Project Gutenberg database as of April, 16 2006 (available at https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#English). Project Gutenberg currently contains the full text of over 70,000 books. Because the database contains books whose copyright has expired, the typical book in the database was published before 1923. This ensures that we exclude frequently used words that reflect historical language use relevant to the period of analysis. The results are robust to excluding only 10,000 or all 36,663 frequently used words reported in Project Gutenberg (Table D.3 and Table D.4). For the main results, we do not remove all frequently used words because words such as quantum (on position 17,132) may have existed earlier but gained new significance in scientific contexts following their use in research publications. For further details on the novel scientific words measure, see Iaria et al. (2018).

Table 2.5.5: Socio-Economic Background and Novelty

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00089*	-0.00101**	-0.00100**	-0.00073*	-0.00090**	-0.00090**
	(0.00048)	(0.00047)	(0.00048)	(0.00043)	(0.00044)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.301	0.301	0.301	-0.002	-0.002	-0.002
Panel B: 1914 – 1969						
Parental SES Rank	-0.00076*	-0.00084**	-0.00085**	-0.00074**	-0.00085**	-0.00087**
	(0.00042)	(0.00041)	(0.00042)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.292	0.292	0.292	-0.011	-0.011	-0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (2.5.7). The dependent variable measures the number of publications which introduce at least one novel word and were published in a ± 5-year window around the cohort when scientist i enters the faculty rosters. We exclude the 20211 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

(Table 2.5.5, column 1, significant a the 10% level). The result is similar and becomes significant at the 5% level if we control for university state and discipline fixed effects (Table 2.5.5, columns 2-3). Specifically, scientists whose fathers were at the 75th percentile of the income rank publish around 0.05 fewer papers (around 17% less) with at least one novel word compared to those whose fathers were at the 25th percentile.

The result is robust to standardizing the novel words measure at the level of disciplines and cohorts (Table 2.5.5, columns 4-6) and in the extended sample (Table 2.5.5, panel B).

2.6 Socio-Economic Background and Recognition

In the last part of the paper, we examine the relationship between socio-economic background and recognition by other academics. First, we analyze citations to a scientist's research papers, a widely-used metric for measuring recognition within the scientific community. Next, we investigate Nobel Prize nominations and awards as indicators of recognition for exceptional scientific contributions.

2.6.1 Citations

To estimate the relationship between socio-economic background and citations, we switch to an analysis at the paper level. This approach allows us to abstract from differences in the number of publications by socio-economic background that we have documented in the previous section. The data include all papers linked to at least one author for whom we can measure parental SES ranks. We estimate the following regression:

$$\text{Citations}_p = \gamma \cdot \text{Avg. Parental SES Rank}_p + \mathbf{X}'_{\mathbf{p}}\beta + \epsilon_p \tag{2.6.8}$$

where Citations_p measures the number of citations that paper p received until 2010. To account for differences in citations across disciplines and over time, we standardize citations at the level of disciplines and the year of publication.[32] Since the distribution of citations is highly skewed and contains outliers,[33] we also estimate results where we winsorize citation counts at the 99th percentile of the discipline and year of publication-specific distribution (Columns 6-10 of Table 2.6.6). Avg. Parental SES Rank_p measures the average SES rank of the fathers (ranging from 0 to 100) of all authors of paper p that we can link to a childhood census. $\mathbf{X_i}$ are controls for the characteristics of the paper. Whenever we measure characteristics at the author level, we aggregate them for all authors of paper p that we can link to a childhood census.[34] As the parental SES rank is based on the father's occupation, childhood state, and birth year, we cluster standard errors at the level of the author team's fathers' occupations, childhood states, and birth years to account for potential correlations of regression residuals.

We find that papers authored by teams from higher socio-economic backgrounds receive more citations (Table 2.6.6, panel A, column 1, significant at the 5% level). Specifically, papers authored by individuals whose fathers, on average, are ranked at the 25th percentile of the income rank distribution receive approximately 0.05 standard deviations fewer citations compared to papers authored by individuals with fathers ranked at the 75th percentile. For example, in medicine, this translates to a paper receiving 2 to 3.5 (13% of the mean) more citations per paper.

The results remain robust, albeit slightly smaller in magnitude when we include fixed effects for the author team's university state and discipline combination, as well as journal fixed effects. In columns 5 and 10, we add fixed effects for both the total number of authors and the number

[32] To capture the whole distribution of citations for the standardization, we use citations to all papers linked to U.S. scientists in the faculty rosters and not only the citations to papers of U.S. scientists, which we can link to a childhood census.

[33] For example, a 1955 medical paper received as much as 61 standard deviations more citations than the average medical paper in that year.

[34] Specifically, we average continuous variables, i.e. we control for the mean age and the share female of the author team, and create a separate fixed effect for each combination of childhood states as well as university states of the author teams.

of authors for which we observe a parental SES rank. When we account for extreme outliers in citations by winsorizing at the 99th percentile (columns 6-10), we estimate coefficients of similar magnitude which are highly significant.

Table 2.6.6: Socio-Economic Background and Paper-Level Citations

Dependent Variable:	*Standardized Citations*					*Winsorized Std. Citations*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: 1900 – 1956										
Average Parental SES Rank	0.00080**	0.00060*	0.00058*	0.00061*	0.00061*	0.00085***	0.00068***	0.00067***	0.00067***	0.00067***
	(0.00033)	(0.00033)	(0.00033)	(0.00032)	(0.00032)	(0.00026)	(0.00026)	(0.00026)	(0.00024)	(0.00023)
R^2	0.03	0.04	0.04	0.10	0.10	0.03	0.04	0.04	0.13	0.14
Observations	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549
Dependent Variable Mean	0.012	0.012	0.012	0.012	0.012	-0.021	-0.021	-0.021	-0.021	-0.021
Panel B: 1900 – 1969										
Average Parental SES Rank	0.00081***	0.00068**	0.00067**	0.00068**	0.00067**	0.00076***	0.00066***	0.00065***	0.00063***	0.00063***
	(0.00029)	(0.00029)	(0.00029)	(0.00027)	(0.00027)	(0.00022)	(0.00022)	(0.00022)	(0.00020)	(0.00020)
R^2	0.02	0.03	0.03	0.10	0.10	0.02	0.04	0.04	0.14	0.14
Observations	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014
Dependent Variable Mean	0.011	0.011	0.011	0.011	0.011	-0.021	-0.021	-0.021	-0.021	-0.021
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publication Year FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes
Discipline FEs			Yes	Yes	Yes			Yes	Yes	Yes
Journal FEs				Yes	Yes				Yes	Yes
Author Count FEs					Yes					Yes

Notes: The table reports the estimates of Equation (2.6.8). The dependent variable measures the number of citations received by paper p until 2010. We standardize citations at the level of displines and years, to account for differences in citations patterns (columns 1-5), and winsorize standardized citations at the 99th percentile to account for extreme outliers (columns 6-10). The main explanatory variable is the average SES rank of the fathers of all authors of paper p that can be linked to a childhood cenus. We proxy the SES rank of fathers with the percentile in the predicted income distribution the father. Demographic controls include age, age squared and the share of female authors. Standard errors are clustered at the level of the author teams' fathers' occupation, childhood states, and birth years. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

2.6.2 Nobel Prize: Nominations and Awards

Nobel Prize Nominations

Next, we study an alternative measure of scientific recognition that captures whether fellow scientists regard a scientist's body of research deserving for a Nobel Prize nomination (Iaria et al., 2018). During this period, Nobel Prize nominations were made by a select group of elite scientists, making the nominations a marker of peer recognition by the scientific elite. We study this question using the following regression:

$$\mathbb{1}\{\text{Nobel Nomination}_i\} = \theta \cdot \text{Parental SES Rank}_i + \mathbf{X_i'}\beta + \epsilon_i \tag{2.6.9}$$

where $\mathbb{1}\{\text{Nobel Nomination}_i\}$ is an indicator for whether scientist i was ever nominated for a Nobel Prize. As before, Parental SES Rank ranges from 0 to 100 and measures the percentile of the income rank of scientist i's father. $\mathbf{X_i}$ are controls as defined above.

We find that individuals from higher parental SES ranks are more likely to be nominated for a Nobel Prize. Specifically, scientists with fathers at the 75th percentile of the income rank distribution have a 0.06 percentage point (or 50%) higher probability of being nominated compared to scientists with fathers at the 25th percentile (Table 2.6.7, column 1).

The results are robust to controlling for the state of the scientist's university and the discipline (Table 2.6.7, columns 2-3). The results are also robust to controlling for both publications and citations (Table 2.6.7, columns 4-6), indicating that scientists from poorer backgrounds are less likely to be nominated for a Nobel Prize even if they have the same number of publications and citations.

Table 2.6.7: Socio-Economic Background and Nobel Prize Nominations

Dependent Variable:	*Nobel Nomination*					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1900 – 1956						
Parental SES Rank	0.00011***	0.00010**	0.00012***	0.00010**	0.00009**	0.00012***
	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)
R^2	0.01	0.02	0.03	0.08	0.08	0.10
Observations	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	0.012	0.012	0.012	0.012	0.012	0.012
Panel B: 1900 – 1969						
Parental SES Rank	0.00010***	0.00009**	0.00010***	0.00009***	0.00009**	0.00010***
	(0.00003)	(0.00003)	(0.00004)	(0.00003)	(0.00003)	(0.00003)
R^2	0.01	0.02	0.03	0.07	0.07	0.08
Observations	15,521	15,521	15,521	15,521	15,521	15,521
Dependent Variable Mean	0.011	0.011	0.011	0.011	0.011	0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Publication & Citation Controls				Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of equation (2.6.9). The dependent variable is an indicator whether a scientist was ever nominated for a Nobel prize. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist is female. Publication and citation controls are a scientist's standardized total publication and citation counts. We standardize publication and citation counts to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the scientist. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Nobel Prize Awards

We also investigate the relationship between the parental income rank and the probability of *winning* a Nobel Prize. We estimate a variant of Equation (2.6.9) with an indicator for winning the Nobel Prize as the dependent variable. We find that scientists from higher parental SES ranks are more likely to win a Nobel Prize (Table 2.6.8). Although some coefficients are not statistically significant, the point estimates remain largely consistent across specifications, regardless of the fixed effects and controls included in the regression. Specifically, scientists with fathers at the 75th percentile of the income rank distribution have a 0.015 percentage point (or 50%) higher probability of winning a Nobel Prize compared to scientists with fathers at the 25th percentile (Table 2.6.7). This finding is robust to controlling for the scientist's publication and citation record.

Overall, these results suggest that socio-economic background plays a significant role in shaping

Table 2.6.8: Socio-Economic Background and Nobel Prize Awards

Dependent Variable:	*Nobel Award*					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1900 – 1956						
Parental SES Rank	0.00003	0.00002	0.00003*	0.00002	0.00002	0.00003*
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
R^2	0.01	0.01	0.02	0.03	0.03	0.04
Observations	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	0.003	0.003	0.003	0.003	0.003	0.003
Panel B: 1900 – 1969						
Parental SES Rank	0.00003*	0.00003*	0.00004**	0.00003*	0.00003*	0.00003**
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
R^2	0.01	0.01	0.02	0.02	0.02	0.03
Observations	15,521	15,521	15,521	15,521	15,521	15,521
Dependent Variable Mean	0.002	0.002	0.002	0.002	0.002	0.002
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Publication & Citation Controls				Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports estimates of a variant of equation (2.6.9). The dependent variable is an indicator whether a scientist was awarded the Nobel prize. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist is female. Publication and citation controls are a scientist's standardized total publication and citation counts. We standardize publication and citation counts to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the scientist. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

peer recognition, as measured by Nobel Prize nominations and awards, with scientists from less privileged backgrounds receiving disproportionately less recognition from the scientific elite.

2.7 Conclusion

This paper examines the role of socio-economic background in shaping the careers of academics and their research output. We show that people from higher socio-economic backgrounds are more likely to become academics and that there is large heterogeneity in representation at the level of universities and disciplines. Further, we find that father's occupation is systematically related to the choice of discipline. Once in academia, socio-economic background is not related to the number of publications, on average, but scientists from lower socio-economic backgrounds are more likely to not publish at all as well as are more likely to have outstanding publication records, making them somewhat riskier hires. The results on novel words suggest that they are somewhat more likely to pursue research agendas off the beaten path which may result in scientific breakthroughs but also in a higher failure rate. We also find evidence that scientists from lower socio-economic backgrounds

receive less recognition by the scientific community, as measured by citations and Nobel Prize nominations and awards. Overall, the paper highlights the importance of understanding the role of socio-economic background in shaping the academic workforce and the creation of new knowledge.

Appendix to Chapter 2

- Appendix 2.A provides further details on the construction of the data.
- Appendix 2.B reports robustness checks and additional findings related to Section 2.3.
- Appendix 2.C reports robustness checks and additional findings related to Section 2.4.
- Appendix 2.D reports robustness checks and additional findings related to Section 2.5.

2.A Appendix: Additional Details on Data

2.A.1 Constructing Parental SES Ranks – Details

As described in the main paper, we use the 1940 census to predict income. We use interactions of fathers' occupation and home state to predict fathers' income for all childhood censuses (see Equation 2.2.1). For some census years and occupations, this approach faces two issues:

1. Rare occupations
2. Changing occupation coding

To overcome these issues, we adjust the income prediction for fathers in affected occupations.

Rare Occupations For a few occupations and states, the number of individuals in certain occupation by state cells in 1940 is low, potentially leading to inaccurate predictions for affected Occupation × State FEs. For example, only four working age male actors reported their income in the 1940s census in Montana, and only one in Wyoming. We thus adjust the income prediction for occupation × state cells with less than 10 observations, by estimating the following regression to predict income:

$$\begin{aligned}\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation}_i \times \text{Region FE} + \beta_2 \text{State FE} \\ + \beta_3 \text{Age}_i + \beta_4 \text{Age}_i^2 + \beta_5 \text{Race}_i + \epsilon_i \end{aligned} \tag{2.A.1}$$

I.e., instead of interacting occupations with states, we interact them with census regions, and estimate a separate state fixed effect.

For even rarer occupations, i.e., those with less than 10 observations in a certain occupation by census *region* cell, we adjust our prediction further:

$$\begin{aligned}\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation}_i + \beta_2 \text{State FE} \\ + \beta_3 \text{Age}_i + \beta_4 \text{Age}_i^2 + \beta_5 \text{Race}_i + \epsilon_i \end{aligned} \tag{2.A.2}$$

Rather than estimating region-specific occupational wage profiles, we now base our prediction on national averages. Only two occupation by region cells are subject to this adjustment: Milliners and Loom Fixers, both in the Mountain Division.

Changing Occupation Coding The Census Bureau has sometimes changed the codes corresponding to specific occupations. For example, the code for actors (and actresses) was 13 from 1850 to 1900, 828 in 1910 and 1920, 192 in 1930, 020 in 1940 and 001 in 1950. To ease comparability

across census years, all earlier census occupation codings were also coded into the 1950s classification scheme by IPUMS (IPUMS, 2024b). We exclusively use the integrated 1950 occupation classification in this paper.

The Census Bureau harmonization process implies that some 1950 occupation codes are present in earlier census years, but not in 1940. For example, the 1950 occupation classification includes codes for "mining engineers" and for "metallurgical engineers", whereas the 1940 occupation classification pools the two engineering fields. In contrast the 1930 and 1920 censuses contain a separate occupation code for "mining engineers."

To address the issue of occupation codes that are aggregated for the 1940 census but disaggregated for earlier censuses, we predict fathers' income via the following regression:

$$\begin{aligned}\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation Group}_i \times \text{State FE} \\ + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i^2 + \beta_4 \text{Race}_i + \epsilon_i,\end{aligned} \tag{2.A.3}$$

where an Occupation Group is the broad one-digit occupational category of an occupation.[35]

Note, that this issue only affects 1.9 % of academics in our data.

2.A.2 Constructing Comparison Group Samples for Other Professions

To compare representation among academics to other professions, we construct samples for lawyers and judges, physicians and surgeons, and teachers, from U.S. Censuses. We proceed in three steps:

1. From each available full-count census corresponding to the coverage period of the World of Academia Database (1900-1950), we extract all observations with occupation code 55 (Lawyers & Judges), 75 (Physicians & Surgeons) and 93 (Teachers).[36]

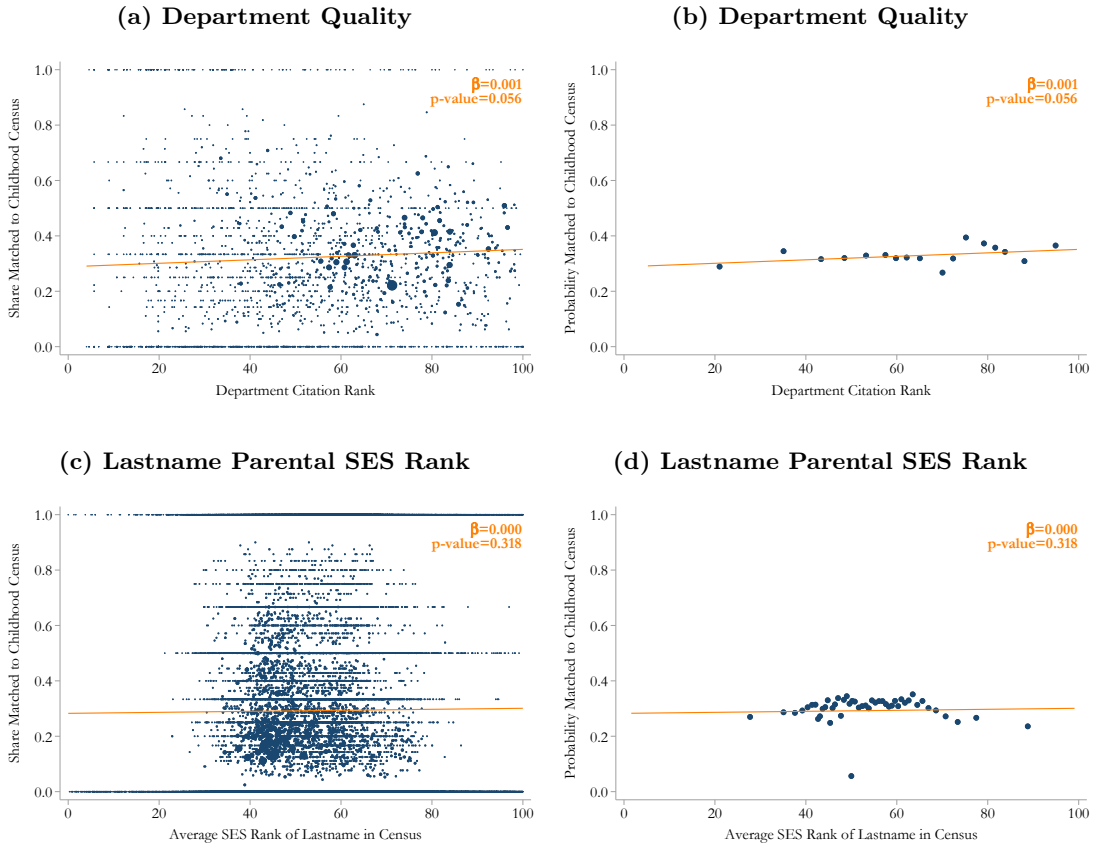
2. We use the Census Linking Project to de-duplicate individuals who appear in multiple censuses and keep only one observation per individual.

3. We then link these observations to their childhood census and construct parental SES ranks as described in Appendix 2.2.2.

[35] Professional, Technical; Farmers; Managers, Officials, and Proprietors; Clerical and Kindred workers; Sales workers; Craftsmen; Operatives; Service workers (private household); Service workers (not household); Farm Laborers; Laborers (non-farm). See IPUMS (2024a).

[36] As discussed in the main text, some academics are not listed as professors but, e.g., as lawyers or surgeons in the U.S. census, we therefore remove all matched academics from this sample.

Figure 2.A.1: Extended Sample 1900-1969: Correlation of Linking Rates With Department Quality and Lastname Parental SES Rank



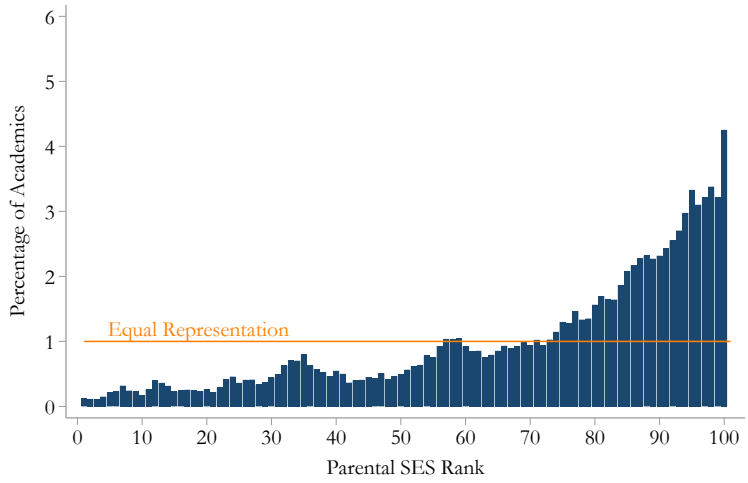
Notes: Panel (a) shows the correlation between a department's citation rank and the probability of linking a scientist to a childhood census for the extended sample (1900-1969). Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's SES Rank based on the entire U.S. census and the probability of linking an academic to a childhood census for the extended sample (1900-1969). Panel (d) shows a binned scatter plot of the same relationship. Bins are chosen according to Cattaneo et al. (2024).

2.B Socio-Economic Background and the Probability of Becoming an Academic: Additional Results

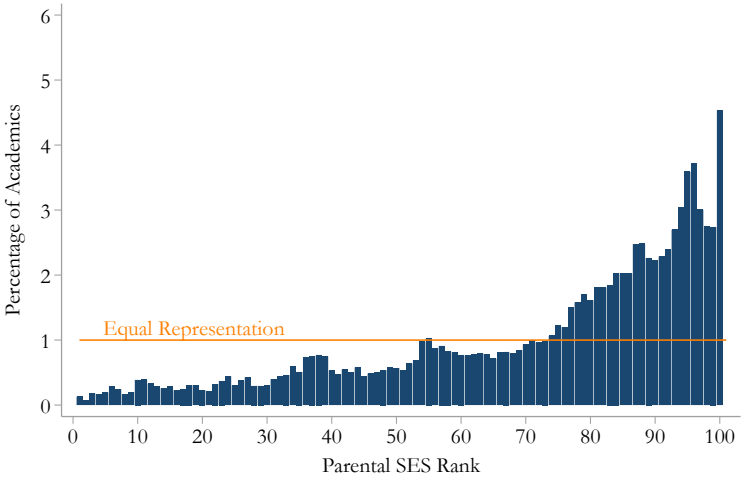
Representation of Academics by Socio-Economic Background

Figure B.1: Representation by Socio-Economic Background, Excluding Children of Professors

(a) With Regional Variation



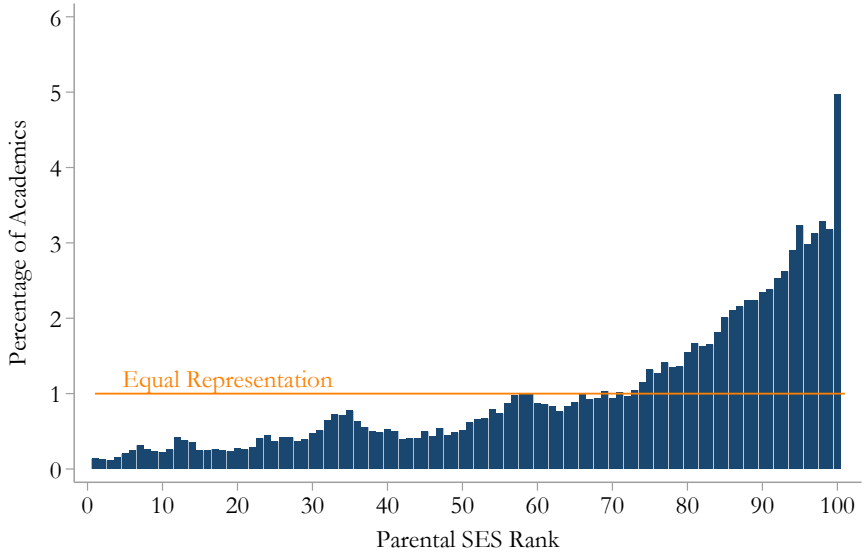
(b) Without Regional Variation



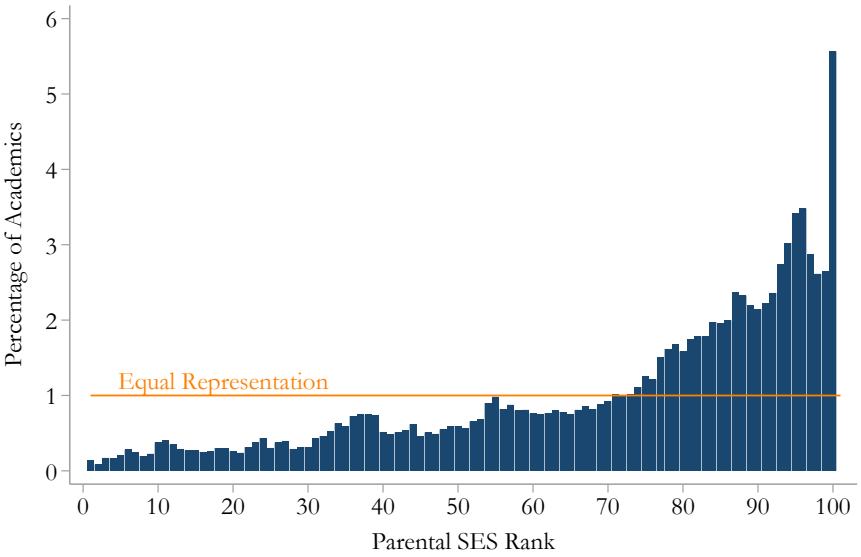
Notes: The figure shows the representation of academics based on their socio-economic background, excluding academics who are children of professors. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. The horizontal line represents a hypothetical equal representation from all income ranks.

Figure B.2: Extended Sample 1900-1969: Representation by Socio-Economic Background

(a) With Regional Variation



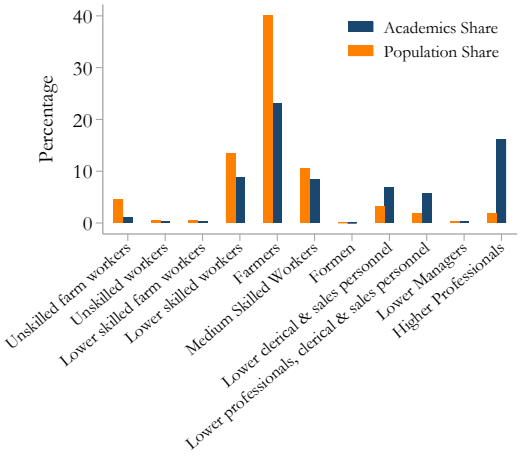
(b) Without Regional Variation



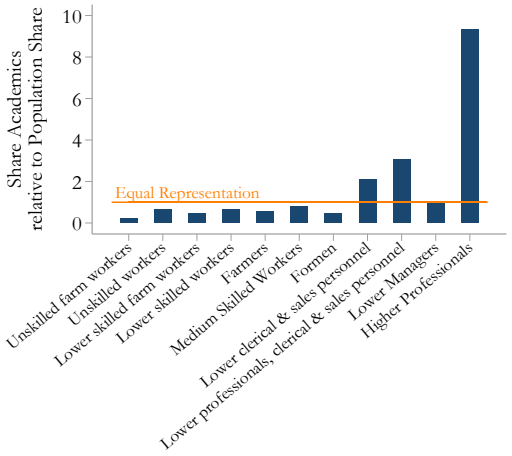
Notes: The figure shows the representation of academics based on their socio-economic background for the extended sample (1900-1969). We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. The horizontal line represents a hypothetical equal representation from all income ranks.

Figure B.3: Representation by Socio-Economic Background, Alternative Measures of SES: HISCLASS

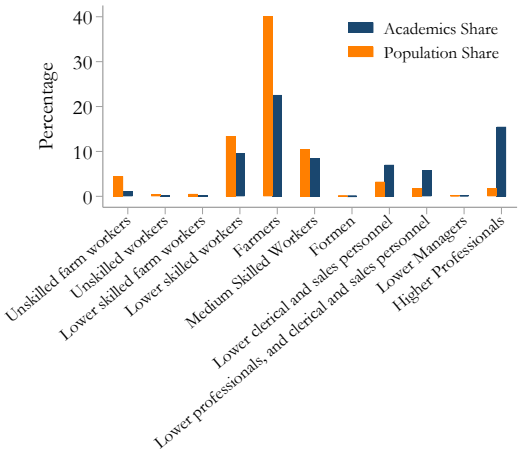
(a) Main Sample 1900-1956



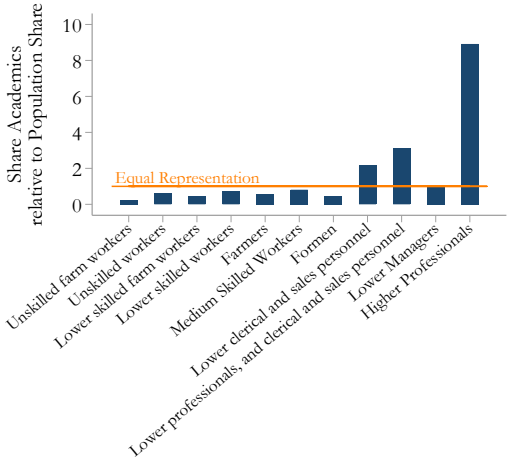
(b) Main Sample 1900-1956



(c) Extended Sample 1900-1969



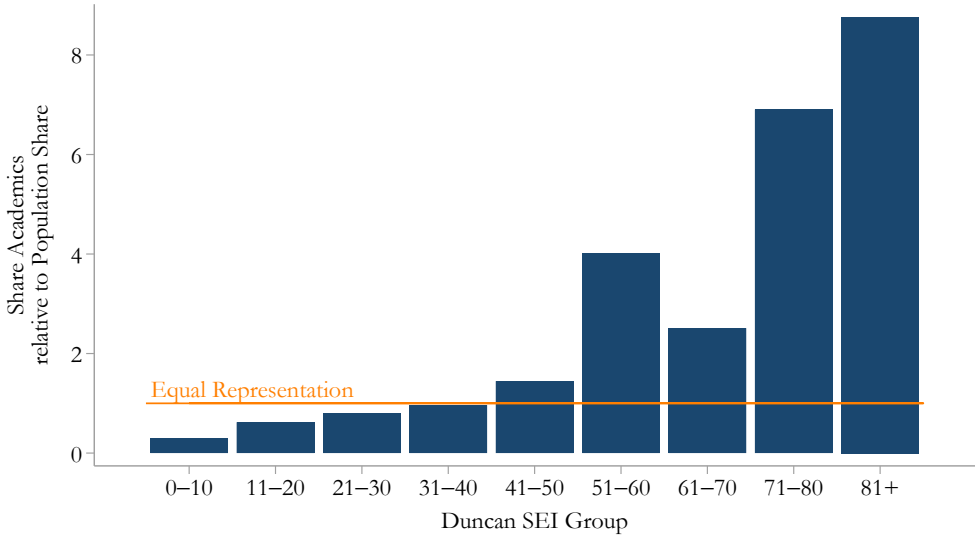
(d) Extended Sample 1900-1969



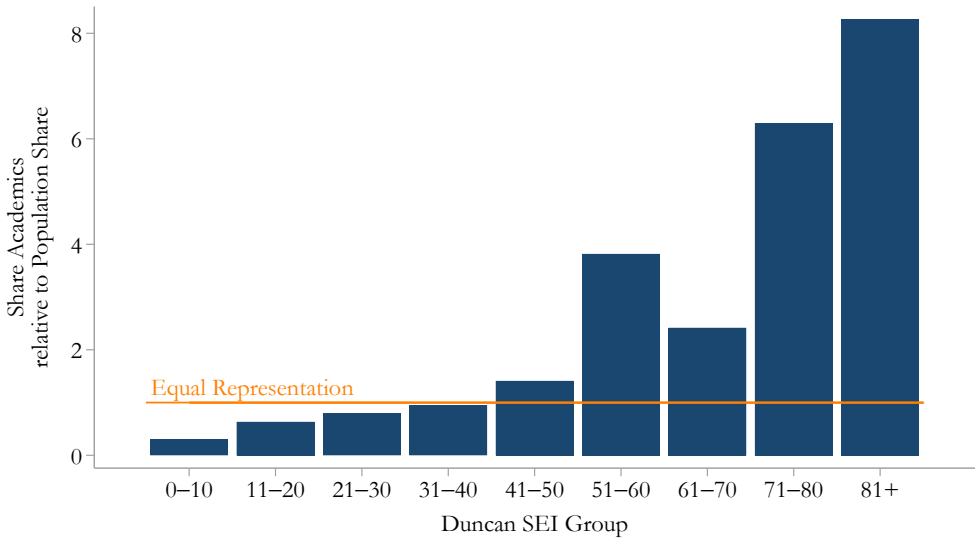
Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with HISCLASS, a measure of the social standing of a father's occupation (van Leeuwen and Maas, 2011). In panels a) and c), the orange bars indicate the share of individuals from a particular HISCLASS in the census. Compared to the census, academics are disproportionately children of fathers in higher status occupations (higher professionals). Panels b) and d) show the share of academics from a HISCLASS relative to the share of the population from the same HISCLASS. The horizontal line represents a hypothetical equal representation of these HISCLASS in the population of academics.

Figure B.4: Representation by Socio-Economic Background, Alternative Measures of SES: Duncan Socioeconomic Index

(a) Main Sample 1900-1956



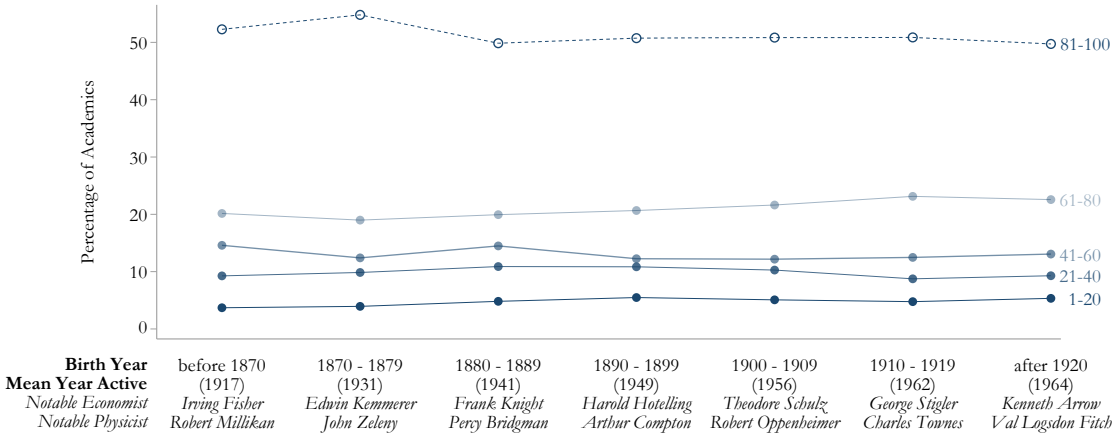
(b) Extended Sample 1900-1969



Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with the Duncan Socioeconomic Index (SEI), a measure of the social standing of a father's occupation. SEI reflects the income level and educational attainment of an occupation in 1950. For details, see IPUMS (2024b). SEI is an ordinal measure of occupational social status with gaps, which we group into 9 categories. For example, the top category, 81+, contains SEI 81-87 (no gaps), 90, 92, 93 and 96. SEI 89 does not exist in the census data of the relevant period. The horizontal line represents a hypothetical equal representation of these SEI categories in the population of academics.

Representation Over Time

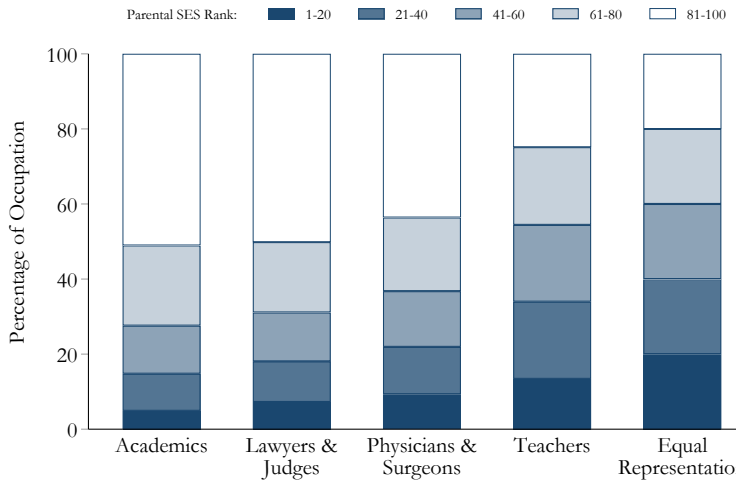
Figure B.5: Extended Sample 1900-1969: Representation by Socio-Economic Background Over Time



Notes: The figure shows the representation of academics based on their socio-economic background over time. Each line represents the percentage of all academics whose fathers are from specific income percentile ranks. For example, the top line indicates the percentage of academics whose fathers are in the top 20 income percentile ranks.

Representation in Academia versus Other Professions

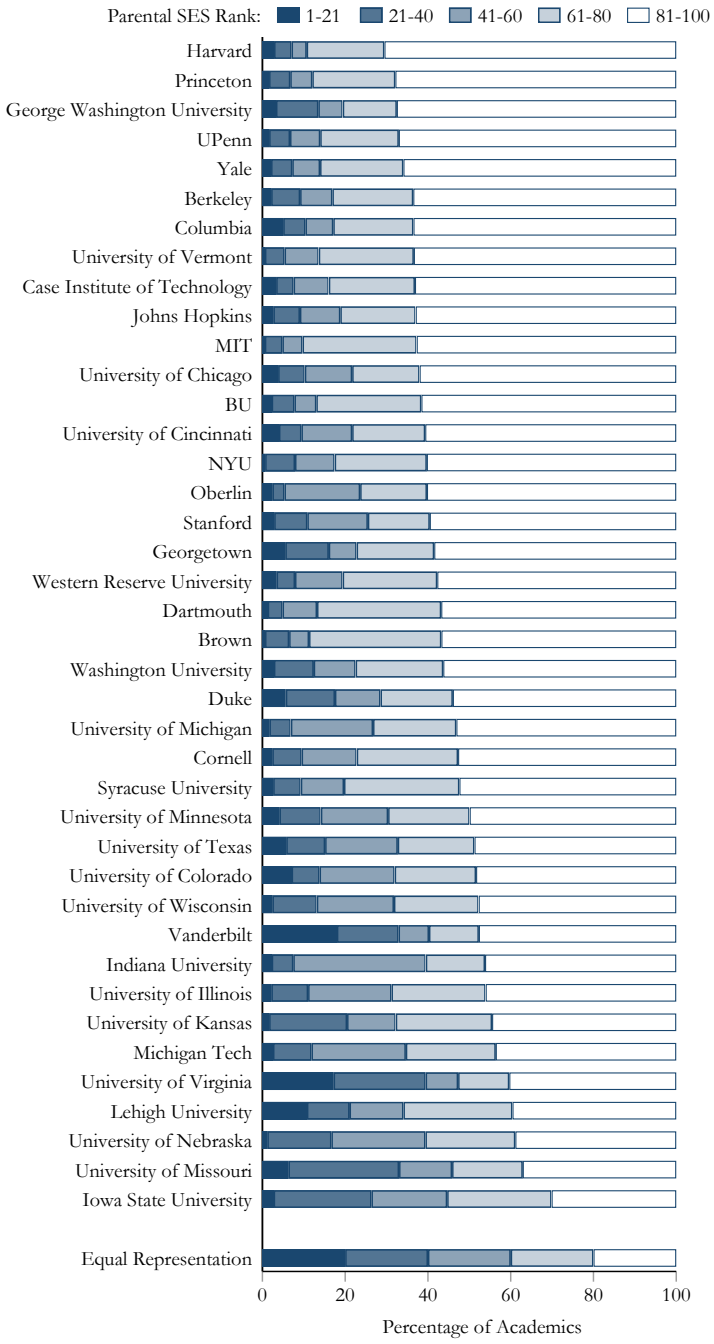
Figure B.6: Extended Sample 1900-1969: Comparison to other Professions



Notes: The figure compares the representation of academics based on their socio-economic background to representation in other professions. The representation in other professions is based on U.S. census samples of lawyers & judges, physicians & surgeons, and teachers that match the sample of academics (see Appendix 2.A.2 for details).

Representation by University: Additional Results

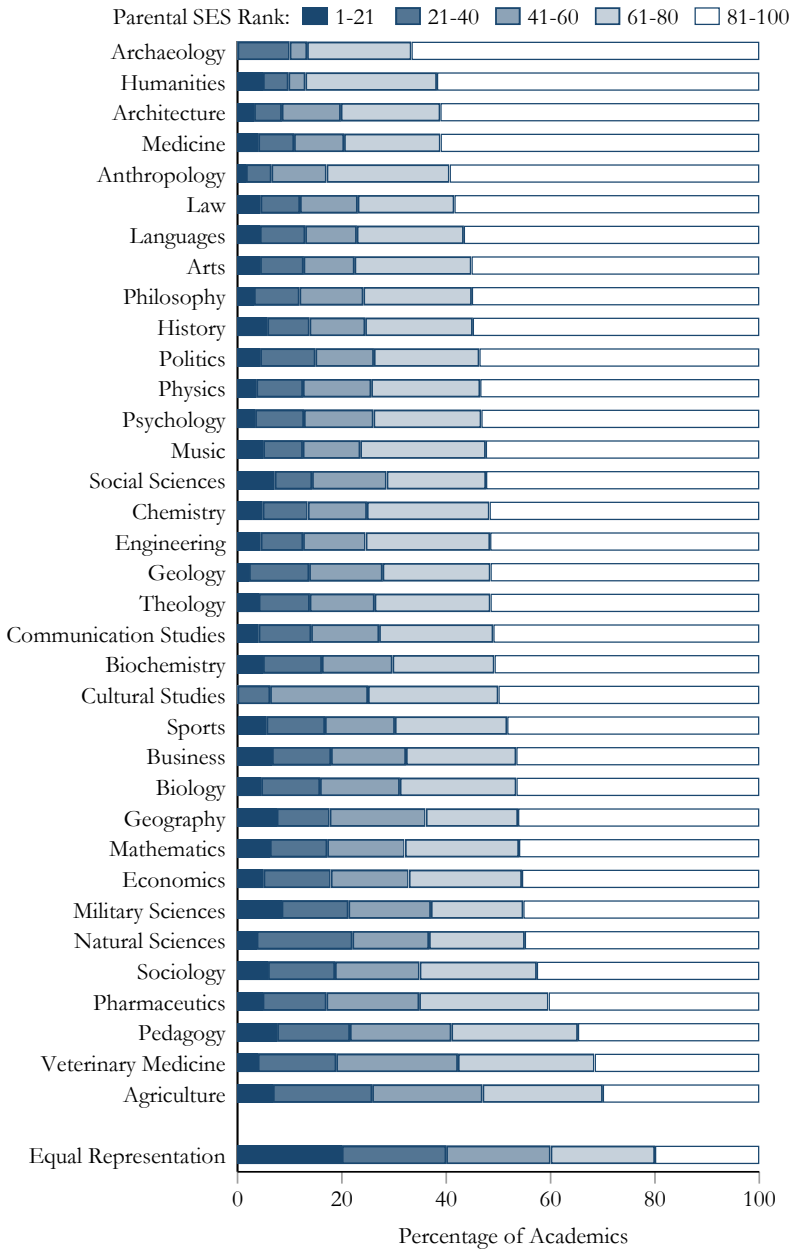
Figure B.7: Extended Sample 1900 - 1969: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

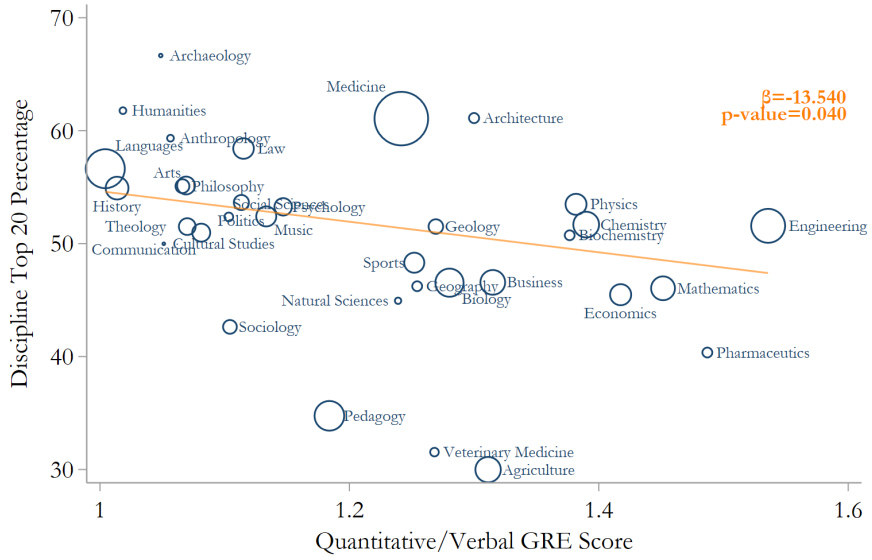
Representation by Discipline: Additional Results

Figure B.8: Extended Sample 1900 - 1969: Representation by Discipline



Notes: The figure shows the representation of academics based on their socio-economic background by academic discipline. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

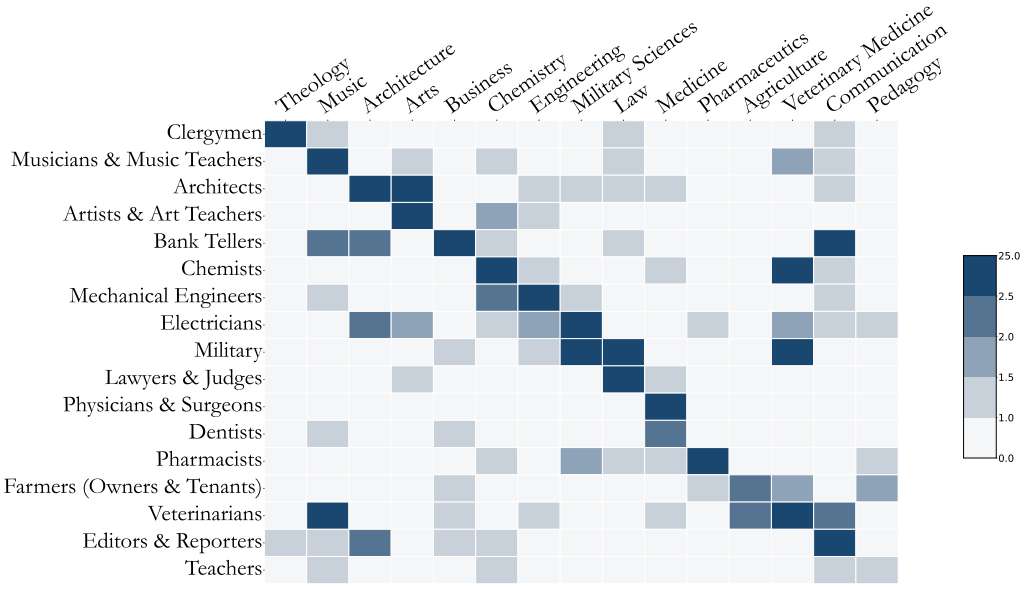
Figure B.9: Extended Sample 1900 - 1969: Discipline Mathematics vs. Language Requirements and Representation



Notes: The figure shows the share of academics from the top quintile of the distribution of socio-economic background by academic discipline in relation to the importance of quantitative relative to verbal skills in the discipline for the extended sample (1900-1969). We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2.2. We proxy the importance of mathematics relative to language skills with the ratio of the average GRE quantitative score to the average GRE verbal reasoning score of test takers intending to pursue a graduate degree in the respective discipline. GRE score data come from ETS (2009), Extended Table 4. The size of the circles indicates the number of academics in the respective discipline in our data.

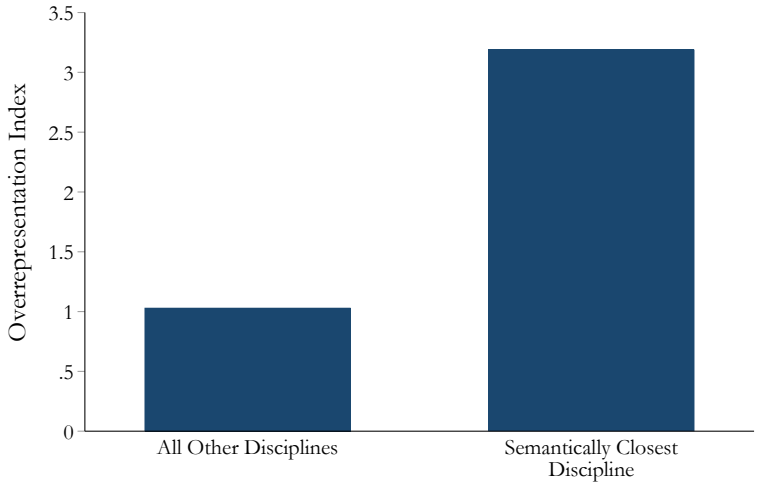
2.C Socio-Economic Background and Discipline Choice: Additional Results

Figure C.1: Extended Sample 1900-1969: Father's Occupation and Discipline Choice



Notes: The figure shows the relationship between father's occupation (rows) and the children's academic discipline choice (columns) for selected father's occupation-discipline pairs. Darker shades indicate more extreme levels of overrepresentation as measured by Equation (2.4.3).

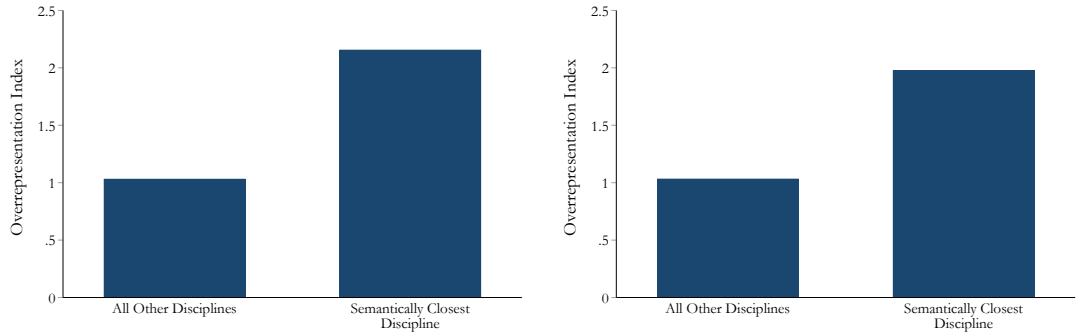
Figure C.2: Extended Sample 1900-1969: Overrepresentation in Semantically Closest Discipline



Notes: The figure shows overrepresentation as measured by equation (2.4.3) in the father's occupation-discipline pair that is semantically closest, e.g., "farmer" and "agriculture" and all other father's occupation-discipline pairs for the main sample. For more details, see appendix 2.4.2 and appendix 2.4.3.

Figure C.3: Robustness – Overrepresentation in Semantically Closest Discipline

(a) One SD Cosine Similarity Cutoff **(b) No Cosine Similarity Cutoff**



Notes: The figure shows overrepresentation as measured by equation 2.4.3 in the father's occupation-discipline pair whose name (e.g., "agriculture") is semantically closest to the text string of the father's occupation (e.g., "farmer") as well as all other father's occupation-discipline pairs. Panel a defines the closest discipline as the discipline that is semantically closest, and the cosine similarity is at least one standard deviation above the mean of all cosine similarities of all father's occupation-discipline pairs. Panel b defines the closest discipline as the discipline that is semantically closest without enforcing a further cutoff on the cosine similarity.

2.D Socio-Economic Background, Scientific Publications, and Novel Scientific Concepts: Additional Results

Table D.1: Publication Percentiles by Discipline and Cohort

		Publication Percentiles					
Discipline	Cohort	*50th*	*70th*	*90th*	*95th*	*97th*	*99th*
Biochemistry							
	1900	1	6	6	6	6	6
	1914	8	13	44	54	58	58
	1925	3.8	9	18	28	30	40
	1938	3	5.5	15.5	22.5	25	57
	1956	4	10	21.5	30	36	50
	1969	5	11	30	41	52	70
Biology							
	1900	0	1	4	7	10	26
	1914	1	2.5	9	13	18	25
	1925	0	2	7	11	13	19
	1938	0	2	6	10	13	20
	1956	1	2	8	11	15	21
	1969	1	4	12	18	22.5	33
Chemistry							
	1900	0	1	6	11	15	58
	1914	1	3	13	19.3	24.5	50.5
	1925	1	3	13	23	27	54
	1938	1	4	16	24	31	63
	1956	1.5	6	21	33	42	64
	1969	2	6	24	39	51	76
Mathematics							
	1900	0	0	3	6	6	13
	1914	0	0	5	8	11	17
	1925	0	0	2	6.5	9	19
	1938	0	0	4	8	12	18.5
	1956	0	0	5	8	11	17
	1969	0	2	9	13	16	24
Medicine							
	1900	0	1	6	9	12	18
	1914	1	3	11	16	21	32
	1925	1	4	13	21	25	42.5
	1938	1	5	15	22	28	44
	1956	2.5	7	20	31	40	59
	1969	3	9	26	40	52	86.9
Physics							
	1900	0	1	7	15	19	37
	1914	1	3	10	12	19	32
	1925	0	2	8	17	24	40
	1938	1	3	10	16	19	30
	1956	1	5	14	21	26	38
	1969	3	9	21	31	39	58

Notes: The table displays the number of publications that place academics in each of these percentiles by discipline and cohort.

Table D.2: Socio-Economic Background and the Distribution of Publications

Dependent Variable:	*Publication Count in Percentile*						
	0 − 50	*> 50 − 70*	*> 70 − 90*	*> 90 − 95*	*> 95 − 97*	*> 97 − 99*	*> 99 − 100*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: 1914 – 1956							
Parental SES Rank	-0.00042**	0.00009	0.00036**	0.00000	0.00006	-0.00002	-0.00007*
	(0.00019)	(0.00014)	(0.00015)	(0.00008)	(0.00005)	(0.00006)	(0.00004)
R^2	0.09	0.03	0.03	0.02	0.02	0.02	0.01
Observations	12,767	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	0.586	0.141	0.180	0.044	0.020	0.019	0.010
Panel B: 1914 – 1969							
Parental SES Rank	-0.00029*	0.00006	0.00028**	0.00010	-0.00007	-0.00002	-0.00006
	(0.00017)	(0.00013)	(0.00014)	(0.00007)	(0.00005)	(0.00005)	(0.00004)
R^2	0.08	0.03	0.03	0.02	0.01	0.01	0.02
Observations	15,521	15,521	15,521	15,521	15,521	15,521	15,521
Dependent Variable Mean	0.557	0.168	0.185	0.045	0.017	0.019	0.008
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Discipline FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports the estimates of eq. (2.5.6). The dependent variable is an indicator whether an academics publication count falls into a certain range of publication percentiles. Publication counts are an academic's total number of publications that were published in a ± 5-year window around the cohort when academic i enters the faculty rosters. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table D.3: Socio-Economic Background and Novelty: Excluding the 10,000 Most Common Words

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00084*	-0.00097**	-0.00096**	-0.00068	-0.00085*	-0.00084*
	(0.00048)	(0.00048)	(0.00048)	(0.00043)	(0.00043)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.305	0.305	0.305	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00073*	-0.00082**	-0.00082**	-0.00070*	-0.00081**	-0.00082**
	(0.00042)	(0.00042)	(0.00042)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.295	0.295	0.295	-0.011	-0.011	-0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (2.5.7). The dependent variable measures the number of publications which introduce at least one novel word and were published in a ± 5-year window around the cohort when academic i enters the faculty rosters. We exclude the 10,000 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table D.4: Socio-Economic Background and Novelty: Excluding the 36,663 Most Common Words

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00094**	-0.00107**	-0.00106**	-0.00081*	-0.00098**	-0.00099**
	(0.00048)	(0.00047)	(0.00047)	(0.00043)	(0.00044)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.296	0.296	0.296	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00080*	-0.00088**	-0.00088**	-0.00079**	-0.00090**	-0.00092**
	(0.00042)	(0.00041)	(0.00041)	(0.00038)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.287	0.287	0.287	-0.011	-0.011	-0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (2.5.7). The dependent variable measures the number of publications which introduce at least one novel word and were published in a $\pm$ 5-year window around the cohort when academic i enters the faculty rosters. We exclude the 36,663 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Chapter 3

Opinions About Facts: Partisan Asymmetries in Economic Assessments

---

with Till Stowasser

This chapter is based on research conducted as part of my master thesis of the same name (submitted in 2019).

3.1 Introduction

Individuals disagree. Yet, in standard economic theory, when signals are public and informative, posterior beliefs converge even if prior beliefs differ. However, systematic partisan differences in beliefs about easily verifiable facts such as – to name a few prominent examples – anthropogenic climate change, effectiveness of gun control, or economic conditions, have consistently been documented. These observations call the assumption that beliefs and preferences can be neatly separated into question. Motivated-cognition theory explains disagreement in terms of preferences over beliefs: Individuals want to "consume" different beliefs.

Motivated cognition has been studied theoretically and documented experimentally. Yet, its relevance for polarization in the field has not been thoroughly investigated. We provide first evidence from observational data that motivated cognition can account for partisan disagreement on facts. Building on a theoretical model of motivated cognition, and using high-frequency survey data around the 2016 U.S. presidential election, we show that Republican perceptions of economic conditions in particular, respond too strongly to be in line with the expected reaction to the actual economic upturn after the unforeseen Republican win. True changes in economic conditions alone cannot explain this reaction. For Democrats, the evidence is less conclusive, but still suggestive of motivated beliefs. Regarding channels, we find tentative evidence in favor of selective interpretation of economic information.

We adapt the theory of motivated cognition to our setting, assuming that partisans derive utility from the belief that their – and only their – party is a competent manager of the economy. Partisans can either ignore or correctly encode information, whichever is conducive to their desired belief. This leads to three predictions: First, partisans disagree. Second, disagreement is sustained by selectively interpreting congenial information. Third, since incentives to distort beliefs change whenever there are changes in political power, there is excess belief movement around power shifts.

In our specific context of an election during an economic upturn, the model predicts a switch to excessively positive perceptions for now-incumbent partisans, who over-weight positive signals, and that perceptions stay at pre-election levels for now-opposition partisans, who disregard positive signals.

Our predictions are supported by the data. We use high-frequency survey data from the United States about the current condition of the economy as our measure for economic perceptions.

The Republican win in 2016 constitutes an unexpected[1] shift in power that allows us to identify structural breaks in perception formation. Correlational evidence from a split-sample ordinary least squares specification already indicates that Republicans judge the economy more favorably

[1]The election was not only close, but a Clinton win was predicted throughout. For a discussion, see Wright and Wright (2018).

during the Republican Presidency than during the Democratic Presidency, even when controlling for economic conditions. The reverse is true for Democrats. Furthermore, increases in the Dow Jones – as a proxy for the general economic climate – are only positively correlated with Republican perceptions of current economic conditions *after* the 2016 election. This suggests that partisans sustain motivated beliefs by selectively leaning on congenial facts.

Since the time period under consideration coincided with a secular economic boom, ordinary least squares cannot disentangle how much of the increase, respectively decrease, in Republican and Democratic perceptions is due to true changes in economic conditions or to motivated cognitive processing of this information. Also, partisans might differ in what indicators they consider relevant for the economy. Our empirical approach needs to be able to account for different partisan mental models of the economy, without us pre-specifying these mental models. Hence, we propose a novel, machine-learning based approach to determine counter-factual economic perceptions: The synthetic belief. We assume that partisan mental models of the economy are constant in the short term, and that perceptions are some function of observable information on the economy. We use LASSO to estimate this function from pre-election data. With post-election economic conditions as an input, this function gives us an estimate of perceptions how they should have been, if only economic conditions had changed, but not political power.

We find that on the day immediately following the election, Republican perceptions already are markedly more positive than their synthetic counterparts. This difference also increases considerably over time, indicating that partisans distort their perceptions of economic conditions into a more desirable direction, suggesting motivated beliefs. There is less pronounced downward divergence for Democratic perceptions. Recall that our model predicts no belief movement for opposition partisans during an economic boom. Still, we would have expected *synthetic beliefs* to increase. Since Democrats' perceptions were remarkably inelastic before the election, we can however not exclude that Democrats' react only to medium- to long-term indicators. By construction, our method cannot incorporate these indicators directly after the election, generating the late increase of synthetic beliefs. At the end of this paper, we briefly discuss alternative explanations for partisan disagreement about economic conditions, such as differential information supply through the media (Hetherington, 1996; Lowry, 2008) or party elites (Bisgaard and Slothuus, 2018), differences in expectations about the future of the economy, and the role of other behavioral biases, such as affect.

To our knowledge, this is the first empirical study in the field of economics explicitly showing the relevance of motivated cognition for partisan disagreement on facts with observational data from the field. Our paper makes three contributions: First, we bring the previously theoretical and experimental literature on motivated cognition to observational data. Second, we provide evidence for the microfoundations of polarization by making the connections between identity and

preferences over beliefs explicit. Third, we make a methodological contribution by providing a novel method, the synthetic belief, for studying cognitive biases in unstructured observational data.

Our research ties into the long tradition of studying partisan gaps in economic perceptions in the political sciences (e.g. Bartels, 2002; Campbell et al., 1960; Schaffner and Roche, 2017). In this literature, partisan disagreement in economic assessments has mostly been attributed to motivated cognition and selective learning. However, there is as yet little empirical evidence to substantiate this hypothesis. An exception is Schaffner and Roche (2017). They show that Republicans over-estimate unemployment whereas Democrats estimate it accurately using the October 2012 Job Reports Announcement as a natural experiment. Since our empirical strategy makes use of both changing political and economic conditions – that is, changing information as well as changing incentives – we can additionally decompose the difference in perceptions into a political and an information-driven effect. Also, we provide tentative evidence for one mechanism how partisans form motivated beliefs, that is, selective interpretation of information.

Within economics, we touch on several literatures. We chiefly contribute to the literature on motivated beliefs, motivated cognition, ideology, and polarization by documenting that empirical patterns of partisan disagreement are consistent with an account of motivated information processing.[2] Akerlof (1989), for example, suggests that motivated beliefs are especially likely to arise in politics, since the negligible probability of affecting the election outcome with one's vote creates little incentive to replace a motivated with an accurate belief. Most closely related to our setting is the theoretical model by Le Yaouanq (2023). He models political polarization on facts as a consequence of different voting preferences. Voters will selectively encode information that is congenial to the belief that their preferred policy option will be enacted. In line with experimental evidence (Eil and Rao, 2011; Zimmermann, 2020), we find suggestive evidence that partisans selectively interpret information.

We also add to the understanding of the determinants of economic beliefs. Using survey and experimental data, Stantcheva (2020) and Alesina et al. (2020) provide evidence for partisan differences in perceptions of facts about economic policy. These differences are reflected in the support for different policies. Our results indicate that partisanship enters at an even more fundamental level. It affects perceptions of economic conditions without reference to any specific policy recommendation. This has, for example, repercussions on economic voting models, for which perceived economic conditions are a key ingredient (Duch and Stevenson, 2011; Fiorina, 1981; Healy and Malhotra, 2013). Finally, we contribute to the study of polarization in economics (e.g., Boxell et al. (2017, 2024); Gentzkow et al. (2019); Draca and Schwarz (2024)) by studying the demand side of polarization: preferences over beliefs.

[2]For a survey of the literature on preferences over beliefs and their consequences, see Bénabou and Tirole (2016).

3.2 Main Hypothesis and Testable Predictions

Partisan disagreement in economic assessments is a long-documented feature of the American political landscape.[3] When information is freely available, Bayesian updating would however predict that – even if individuals hold partisan priors – these asymmetries disappear over time. We thus hypothesize that they are a result of *motivated cognition*: We propose that partisans derive utility from the belief that their party is a competent manager of the economy. Consequently, their preferences over beliefs about economic conditions are state-dependent: Good economic conditions are preferable over bad economic conditions whenever their party is incumbent, and vice versa when it is in opposition. In order to maintain this belief, they exercise selective interpretation of information. Thus, congenial facts – that is, information which indicates that the preferred party is doing well or the non-preferred party is doing badly – enter perceptions at a higher rate than contrary facts.

Theoretical Framework To fix ideas, we illustrate our predictions with a stylized model of motivated reasoning. The purpose of this model is not to provide a full model of motivated cognition, but to provide a simple, tractable and intuitive framework for our main predictions.[4] To this end, our assumptions on the utility function and the information structure are, without loss of generality, unnecessarily stark, quite simply to reduce the complexity of the model. The predictions of the model also hold for less stringent assumptions.

The economy is populated by partisans who are identical in all respects except for their status as supporters of either the incumbent party ($\rho = I$) or an opposition party ($\rho = O$). A partisan[5] has state-dependent preferences $U^{\rho}(e)$ about their economic perceptions e. The economy has two (unobservable) states $S \in \{G; B\}$, good G and bad B. The economy is in the good state with probability $p = P(S = G)$. A perception is the belief that the economy is in the good state, i.e. $e = P(S = G) = p$. The key assumption in this model is that a partisan prefers good economic conditions during incumbency rather than opposition, and bad conditions during opposition rather than incumbency. Although this assumption is sufficient to generate the predictions of our model, we further simplify the preference structure in our model to the case where partisans are sore losers: $U'^{I}(p) > 0$ and $U'^{O}(p) < 0$, meaning that incumbent partisans' utility increases in the belief that the economy is in the good state (increasing p), whereas opposition partisans' utility decreases. p is unknown to partisans. We denote their prior belief about p by p_0, where $p_0 \in (0, 1)$. Partisans receive a signal s about the state of the economy, where $s \in \{G; B\}$. s is informative about the

[3] See, for example, Campbell et al. (1960); Bartels (2002); Schaffner and Roche (2017).

[4] For more thorough models of motivated cognition and political behavior, see e.g. Bénabou and Tirole (2011, 2006) or Le Yaouanq (2023).

[5] For readability, we do not include individual indices in the following.

true state of the economy with probability q, where $q > 0.5$. We denote the realization of signal s by $\hat{s}$. For the formation of motivated beliefs, a crucial requirement is that partisans are able to distort their information processing. We assume that upon receiving signal $\hat{s}$, partisans can choose to either encode signal $\hat{s}$ correctly or disregard the signal and instead encode $\hat{s} = \emptyset$ at no cost. This assumption is similar to the selective recall technology employed by, among others, Bénabou and Tirole (2002, 2011) and Le Yaouanq (2023). If s is encoded correctly, posterior beliefs $p_1^{\hat{s}} = P(S = G|s = \hat{s})$ will be determined by Bayes' Rule.[6] Otherwise, beliefs remain unchanged at priors: $p_1 = p_0$. After encoding signals, partisans receive utility $U^\rho(e = p_1)$. Belief-utility maximizing partisans will thus encode signals correctly if and only if the utility from correct posteriors is larger than the utility from priors, that is $U^\rho\left(p_1^{\hat{s}} = P(S = G|s = \hat{s})\right) > U^\rho(p_0)$. Since $U^\rho(\cdot)$ is strictly increasing in p for incumbent partisans – and strictly decreasing for opposition partisans – this condition boils down to: Encode $s = \hat{s}$ if $p_1 > p_0$ and $\rho = I$, or $p_1 < p_0$ and $\rho = O$, else, encode $\hat{s} = \emptyset$. This gives rise to four different cases:

Table 3.2.1: Updating Strategies for Incumbent (I) and Opposition Partisans (O)

	$\hat{s} = B$	$\hat{s} = G$
$r = I$	$p_1^B < p_0$ encode $\hat{s} = \emptyset$	$p_1^G > p_0$ encode $\hat{s} = G$
$r = O$	$p_1^B < p_0$ encode $\hat{s} = B$	$p_1^G > p_0$ encode $\hat{s} = \emptyset$

Accordingly, incumbent partisans will only update in response to good signals, while opposition partisans update exclusively in response to bad signals, and partisans – even if prior beliefs p_0 are identical – will hold different posterior beliefs p_1. We derive three testable predictions that we view as litmus tests for partisan motivated reasoning:

Prediction 1: Partisan disagreement Incumbent partisans perceive the economy more favorably than opposition partisans. In the model, incumbent partisans' perceptions of the economy are either p_1^G or p_0 and opposition partisans' are either p_1^B or p_0, where $p_1^G > p_0 > p_1^B$.

Prediction 2: Selective relevance of economic information Partisans will only update in response to economic information that is congenial to their desired beliefs. Incumbent partisans' utility increases in $P(S = G)$ while opposition partisans' utility decreases in $P(S = G)$ – and hence increases in $P(S = B)$. As seen in Table 3.2.1, incumbent partisans only update if $\hat{s} = G$ and opposition partisans if $\hat{s} = B$.

[6] $p_1^G = \frac{qp_0}{qp_0+(1-q)(1-p_0)}$ and $p_1^B = \frac{(1-q)p_0}{(1-q)p_0+(1-p_0)q}$.

Prediction 3: Disproportionate perception changes at power shifts When power changes, incumbents become the opposition and vice versa. This changes partisans' objective functions, and hence the signals they respond to. Since partisans only update in one direction, changing that direction and thus incorporating an opposite signal will move their beliefs more strongly than warranted by the informative content of the signal. We expect to detect these patterns in the data: Incumbent partisans assess the economy more favorably than opposition partisans, with belief movement only in response to congenial signals, and sharp changes in perceptions when power shifts from the incumbent to the opposition.

Let us situate our predictions within the experimental and theoretical literature about motivated cognition: Previous research has shown that individuals desire to believe themselves to be moral, intelligent, and attractive (Bénabou and Tirole, 2011; Eil and Rao, 2011; Gino et al., 2016; Zimmermann, 2020). We posit that this extends to social identity, such as partisanship. Westwood et al. (2018) provide evidence that, in the United States, party affiliation is more central to individual identity than class or race – both of which have been shown to have considerable impact on economic choice (Akerlof and Kranton, 2010). Politicians are often evaluated in terms of the performance of the economy (Fiorina, 1981; Wolfers, 2007). In this line of reasoning, the beliefs about a social group, such as a party, that one is affiliated with, can enter an individual's utility function. Similar to the incentive to see oneself as moral and intelligent, partisans might wish to see their party as more competent, also as regards the management of the economy. In this case, there is a clear individual incentive to distort signals to gain a more rosy view of one's own party – whether that is by inflating performance of the party in question, or devaluing an opposing party's performance. The literature suggests several possible channels for belief manipulation, such as information avoidance (Golman et al., 2017; Oster et al., 2013), or selective recall (Bénabou and Tirole, 2002, 2011; Eil and Rao, 2011; Zimmermann, 2020). Both lead to congenial information being favored over uncongenial information, either at the stage of information acquisition or information processing. Because we cannot differentiate between them in our setting, we stay agnostic towards which exact factor creates motivated beliefs in the empirical section. However, both arise only if partisans have preferences over beliefs, that is, if motivated cognition is present.[7] The next section describes the data which we use to test our predictions.

3.3 Data

For our analysis, we combine high-frequency online opinion-polling data with a range of economic time-series data.

[7] In the course of this paper, we will use the terms "selective interpretation" and "selective relevance" interchangeably to refer to both.

3.3.1 Survey Data

Survey data comes from Civiqs.com, a provider of online opinion polling and survey services. We use both a repeated cross section of individual-level responses and a time series of aggregate data from Civiqs' tracking polls. These polls are conducted on a daily basis and include a range of political and economic questions, of which we use *National Economy, Current Condition* as a measure of economic perceptions. The exact wording of the question is "How would you rate the condition of the national economy right now?". Possible answers are "very good", "fairly good", "fairly bad", "very bad", and "unsure". We re-code these answers to numerical values to facilitate analysis. "Very good", "fairly good", "fairly bad", "very bad" are coded to be equal to 3, 2, 1, and 0 respectively. Answers "unsure" are coded as missing.[8]

Our analysis makes use of the entire time series of survey responses sampled and aggregated by partisanship by Civiqs since January 2015. Additionally, we have access to a random sample of 1100 individual person-day observations from May 1st, 2016 to April 29th, 2017. Since we originally planned to conduct a regression-discontinuity in time analysis – which turned out to be underpowered - we sampled more observations closer to the 2016 presidential election.[9] The sample contains roughly 50% of Republicans and Democrats each. It does not contain any Independents. On the individual-respondent level we observe perceptions, which are at the center of our analysis, as well as a range of self-reported characteristics such as gender, race, age, income bracket, education level, home state, home city, and whether the individual characterizes their home area as rural, urban, or suburban.

Table 3.B.2 provides summary statistics. On average, Democrats view the the condition of the economy more favorably than Republicans during the sampling period. Democrats are also significantly more female, racially diverse, younger, they tend to live in more urban areas, and are slightly better educated than Republicans in our sample.

Survey Design and Sample Selection Civiqs uses list-based sampling to select survey respondents from a representative pool of panelists. They then aggregate responses via a dynamic Bayesian multiple regression model with post-stratification weights.[10] Both methods aim to correct for underrepresented groups in the panel. Appendix Figures 3.B.2a and 3.B.2b assess representativeness of our sub-sample with respect to party demographics and them overall U.S. population, respectively.

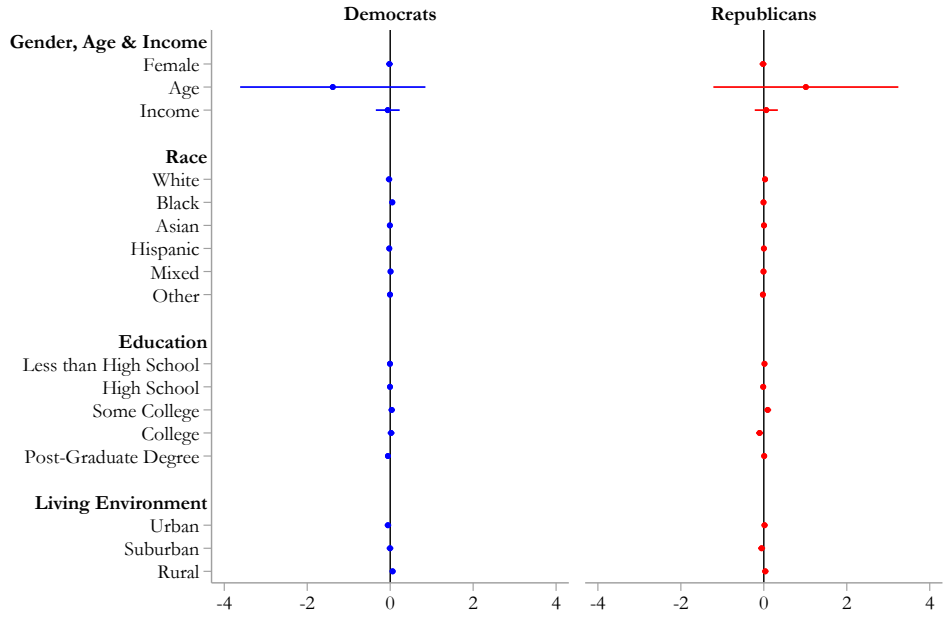
[8] We exclude "unsure" answers because we have no predictions on cognitive uncertainty. In the augmented pre-analysis plan we provide an analysis of "unsure" responses and cannot detect any partisan pattern.

[9] For an illustration, see Appendix Figure 3.B.1. To be exact, we observe the 500 respondents closest to the election threshold on both sides and a random sample of 50 observations for each month between May 2016 and April 2017. For details, we refer the reader to our pre-analysis plan.

[10] For more details on Civiqs' sampling methodology, see https://civiqs.com/methodology/.

A concern is whether different individuals select into answering the survey pre- and post-election. Although we can control for these effects in the individual data, we need to determine if there are selection effects that might bias results derived from the aggregate time-series. For this reason, we check for selection in the individual data. Figure 3.3.1 documents that there are no significant differences between pre- and post-election respondents for each party. Appendix Figure 3.B.3 repeats this exercise with respondents' home states and shows similar results. Taken together, there is no evidence for selection based on observable characteristics.

Figure 3.3.1: Pre-and Post Election Characteristics of Survey Sample



Notes: Difference in pre- and post-election characteristics of partisans. Female is an indicator that is equal to one if the respondent self-identifies as female (1) or male (0). Third gender/non-binary gender is coded as missing due to few observations. Age is a respondent's age, measured in years. Income is reported in units of $25,000. White/Black/Asian/Hispanic/Other are indicator variables that are equal to one if the respondent identifies with a given ethnicity. Ethnicities with fewer than 20 observations are included in "Other". Education variables are indicators that are equal to one if the respondent falls into a given educational category. Urban is an indicator that is equal to one if the respondent lives in an urban area, suburban is an indicator that is equal to one if the respondent lives in a suburban area, and rural is an indicator that is equal to one if the respondent lives in a rural area. Lines indicate 95% confidence intervals.

3.3.2 Economic Data

We use a large set of over 90 economic indicators in the analysis of our aggregate data. These cover macro indices, interest rates, prices and inflation, households income, as well as housing-market,

labor-market, firm, government, trade, and stock-market indicators. When available, we use non-seasonally adjusted data since the survey question explicitly refers to economic conditions at the very moment. For details, see Appendix Table 3.A.1. Since data is reported at different frequencies, and our survey data is daily, we need to interpolate monthly, quarterly, and annual data. Our main empirical specification uses time series that simply keep the value of the indicator constant over the reporting period. We also include lags and growth rates of all variables. In all OLS specifications, we only use the Dow Jones Composite Average as a proxy for economic conditions to reduce multicollinearity and interpolation-induced measurement error, since the Dow is reported daily.

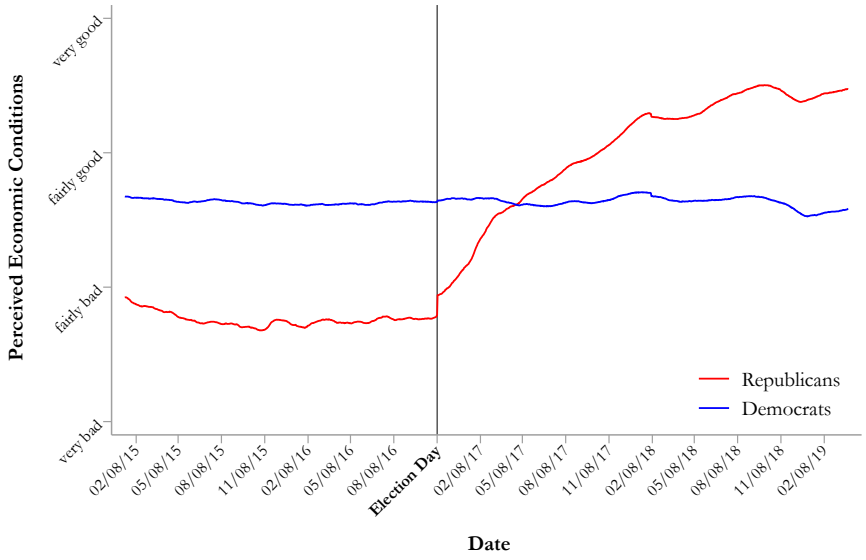
3.4 Descriptive and Regression Evidence

We use the November 8th, 2016 presidential election to test for motivated cognition in partisan perceptions of economic conditions. This particular election has the advantage that the victory of Republican candidate Donald Trump was largely unanticipated (Wright and Wright, 2018). Thus, perceptions should not have adjusted prematurely to a Republican win, which creates a sharp cut-off between the two presidencies. Testing our predictions in U.S. data also limits concerns of endogeneity – that individuals change partisan affiliation according to their momentary beliefs about economic conditions – because U.S. party identification is particularly strong and stable over the life cycle (Green and Palmquist, 1994; Bartels et al., 2011). We start by presenting descriptive and correlational evidence that is indicative of the presence of motivated cognition.

According to Prediction 1, Republican perceptions of the economy should be more positive during a Republican presidency than during a Democratic presidency, while the opposite should be true for Democrats. To get a first picture of the dynamics of partisan perceptions, we plot the entire time series of aggregate perceptions in Figure 3.4.2. Prediction 1 is borne out in the data – Republicans perceive the economy more favorably than Democrats during a Republican presidency, and vice versa during a Democratic presidency. Republican perceptions also change discontinuously at the election threshold, which is in line with Prediction 3. We observe no discontinuities for Democrats. However, recall that our model does not predict any change in beliefs for opposing partisans, if economic conditions improve. In conclusion, while the descriptive evidence is highly indicative of the presence of motivated cognition, it is inconclusive as long as we do not incorporate economic conditions into the analysis.

To test whether there still is a partisan response to the changing political environment once we

Figure 3.4.2: Aggregate Trends in Perceptions



Notes: Partisan aggregate time series from Civiqs of survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" The responses are re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Data covers 01/15/2015 to 03/31/2019.

control for economic conditions, we estimate the following regression using ordinary least squares:

$$E_{it} = \beta_0 + \beta_1 RPresid_t \times Republican_i + \beta_2 RPresid_t + \beta_3 Republican_i + \beta_4 DJCA_t + \mathbf{X'_{it}}\gamma + \epsilon_{it}. \tag{3.4.1}$$

E_{it} is individual i's perception of the economy, $RPresid_t$ is a dummy that indicates a Republican presidency, and $Republican_i$ indicates whether individual i affiliates with the Republican party. $DJCA_t$ is the Dow Jones Composite Average at date t, and $\mathbf{X'_{it}}$ a vector of socio–economic controls. We opted for the DJCA as a proxy for economic conditions since it is a single indicator that is very responsive to changes in economic fundamentals and is reported daily.[11] Controlling for economic conditions with the DJCA allows us to address our first two predictions. Prediction 1 posits that partisans perceive the economy differently, and that this depends on who is in power. For this, perceptions of Republicans and Democrats need to respond to changes in political conditions, and not just to changes in economic conditions. If only the latter were the case, we would not expect to see a significant coefficient for the presidency dummy, and there would not be a case for motivated

[11] Originally, we were also concerned that partisans report their expectations instead of their perceptions. The DJCA would also mitigate this issue if present, since it incorporates expectations of future economic conditions as well. We further discuss the distinction of expectations and perceptions, and why we are not concerned that our survey respondents report the wrong object, in Section 3.6.

cognition. Second, we can develop a first test for Prediction 2: Do economic conditions affect perceptions differentially depending on which party holds power? If Prediction 2 holds, we would expect to see significant coefficients for the interactions of the DJCA and presidency.[12]

Table 3.4.2 shows results for the first exercise. We can exclude that perceptions only reflect economic conditions, since the coefficient on the Republican presidency is significant, even when taking the positive association between the DJCA and perceptions into account. Taking another look at Prediction 1, Table 3.4.2 clearly demonstrates that Democrats perceive the economy as better than Republicans during the Democratic presidency (the coefficient for "Republican" is negative), that the Republican presidency has a small negative effect on Democratic perceptions, and a stronger positive effect on Republican perceptions. This all is in line with Prediction 1. The Dow Jones has a strong positive association with perceptions. Column (2) contains a full set of state fixed effects. Appendix Table 3.B.3 displays all coefficients for sociodemographic controls. Apart from income and one ethnicity dummy – Asian – (both positive) no demographic factor has significant impact on perceptions.

Table 3.4.2: Prediction 1: Partisan Disagreement

Dependent Variable	Perceptions	
	(1)	(2)
Republican Presidency × Republican	0.3419***	0.3628***
	(0.0934)	(0.0952)
Republican Presidency	-0.1558*	-0.1801**
	(0.0864)	(0.0905)
Republican	-1.2148***	-1.2401***
	(0.0694)	(0.0739)
Dow Jones CA	0.0004***	0.0005***
	(0.0001)	(0.0001)
Constant	-0.7832	-0.7858
	(0.8007)	(0.8649)
Demographic Controls	Yes	Yes
State FEs		Yes
Adjusted R^2	0.42	0.42
Observations	874	874

Notes: The table reports results for Equation (3.4.1) testing Prediction 1: Partisan Disagreement irrespective of economic conditions. The dependent variable are perceptions measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?", responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing). Demographic controls contain gender, race, age, income bracket, education level, living environment (rural/urban/suburban). Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

To assess Prediction 2, we test whether economic conditions enter perceptions differently depending on partisanship and presidency. Recall that our model predicts that partisans update in response to positive signals during incumbency, and in response to negative signals during opposition.

[12]Note that we cannot provide parametric evidence for Prediction 3, since a regression-discontinuity-in-time specification, which would allow us to separate short- and long-term effects of the election, and thereby test for disproportionate perception changes at power shifts, is underpowered for our sample.

Because the triple interaction between partisanship, presidency, and economic conditions is difficult to interpret, we estimate the following empirical specification separately for Republicans and Democrats:

$$E_{it} = \beta_0 + \beta_1 RPresid_t \times DJCA_t + \beta_2 RPresid_t + \beta_3 DJCA_t + \mathbf{X_{it}'}\gamma + \epsilon_{it}. \tag{3.4.2}$$

All variables are defined as above. Looking at the results for Republicans in Table 3.4.3, Panel (a), the Dow Jones is positively correlated with perceptions only during the Republican presidency, although it was on an upward trend throughout the entire sample period. As by Prediction 2, positive signals (increases in the DJCA) only enter during incumbency (for Republicans during a Republican presidency). Surprisingly, perceptions are neither associated with economic nor political variables for Democrats (Table 3.4.3, Panel (b)). According to our theoretical model, there should have been a positive association between the Dow Jones and perceptions during the Democratic presidency. A possible explanation for this finding is that, as suggested by Bartels (2002) and others, Republicans and Democrats differ in terms of what economic indicators they consider sufficient statistics for the general economy. Partisan perceptual differences would then arise because of differences in sufficient statistics.[13] This brings us to the limitations of an OLS strategy when testing for motivated cognition in unstructured opinion data. We cannot control for a larger set of potentially relevant economic determinants because of collinearity between economic indicators. Choosing a different set of indicators for Republicans than for Democrats would be highly arbitrary, and vulnerable to specification search. A further limitation of OLS is that it cannot account for feedback effects between our two right-hand side variables. If the election influenced the economy, or the economy the election – both highly plausible – OLS cannot disentangle the two effects. Finally, OLS further highlights our missing counterfactual problem. To determine whether partisans exhibit motivated cognition, we need to know how perceptions would have developed in absence of motivated cognition. However, we lack a clear control group – Republicans are not a good control for Democrats and vice versa, since they are both affected by the election and our theory predicts heterogeneous effects of the election on the two partisan groups. We also cannot simply use any economic indicator as a counterfactual, since we would have to make strong assumptions on how a specific indicator translates into economic perceptions for partisans. This motivates our development of a novel, data-driven empirical strategy: the *Synthetic Belief*.

[13] Hibbs et al. (1982), for example, provide evidence that Democrats emphasize unemployment and Republicans inflation. Disagreement in perceptions arises as a result of pure macroeconomic connection.

Table 3.4.3: Prediction 2: Selective Relevance of Economic Information, by Partisanship

(a) Republicans

Dependent Variable	Perceptions	
	(1)	(2)
Republican Presidency × Dow Jones CA	0.0012**	0.0013**
	(0.0006)	(0.0006)
Republican Presidency	-7.7885**	-8.2444**
	(3.6148)	(3.8514)
Dow Jones CA	-0.0000	-0.0001
	(0.0005)	(0.0006)
Constant	0.5985	0.9881
	(3.4715)	(3.6426)
Demographic Controls	Yes	Yes
State FEs		Yes
Adjusted R^2	0.16	0.14
Observations	432	432

(b) Democrats

Dependent Variable	Condition	
	(1)	(2)
Republican Presidency × Dow Jones CA	0.0004	0.0004
	(0.0005)	(0.0005)
Republican Presidency	-2.3151	-2.5718
	(3.1590)	(3.3171)
Dow Jones CA	-0.0004	-0.0005
	(0.0005)	(0.0005)
Constant	5.2801*	5.5570*
	(2.9796)	(3.2206)
Demographic Controls	Yes	Yes
State FEs		Yes
Adjusted R^2	0.03	0.04
Observations	442	442

Notes: The table reports results for Equation (3.4.2), testing Prediction 2: Partisan Disagreement sustained by selective relevance of economic information. The dependent variable are perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?", responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing). Panel (a) displays results for Republicans, and Panel (b) displays results for Democrats. Demographic controls contain gender, race, age, income bracket, education level, living environment (rural/urban/suburban). Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

3.5 Synthetic Beliefs

As a solution to the missing counterfactual problem, we propose a novel method: *Synthetic beliefs*. Synthetic beliefs predict post-election perceptions of the economy as if the election had only affected perceptions through its indirect impact via economic conditions but not directly via preferences over beliefs. Using this paradigm, we find that Republican perceptions are considerably higher than warranted by previous perception formation, and the perceptions of Democrats are lower. This is in line with what our theoretical model of motivated cognition, specifically Prediction 2 and 3, predicts: Now-incumbent partisans change their perceptions abruptly after the election (Prediction 3), thereby reacting too strongly to positive economic developments (Prediction 2),

whereas now-opposition partisans react too little or not at all (Prediction 2 and 3).

3.5.1 Synthetic Beliefs as Counterfactual Perceptions

The largest challenge for identifying motivated cognition in observational data is to find a benchmark for non-motivated beliefs. Our identifying assumption is that without motivated cognition, there is a stable association between economic perceptions and observable economic information. This association is unaffected by the political environment. Imagine we could write perceptions as a simple function of economic conditions. Without motivated cognition, the political environment only changes the argument of the function, that is, economic conditions, but not the functional form. However, according to our theoretical model, under motivated cognition the political environment changes *how* economic conditions translate into perceptions, that is, it changes the functional form of the perceptions-economic conditions function.

We adapt this exact logic to our estimation strategy: We estimate perceptions as a function of only economic conditions. These are our synthetic beliefs, which are by construction independent of the political environment. They provide benchmarks of how perceptions should have developed in the absence of any changes aside of economic change. For a more theoretical exposition, see Appendix 3.C.

In a nutshell, synthetic-belief analysis proceeds in three steps: In step 1, we use LASSO to determine the set of best predictors for perceptions in the pre-election period. Second, we predict synthetic beliefs for the post-election period using the model selected in step 1. Third, we compare synthetic beliefs to observed perceptions. If the election has no impact on perceptions except through its direct impact on the economy, we expect synthetic beliefs to closely track observed perceptions. However, if there is a political effect over and above the economic effect, synthetic beliefs and perceptions will diverge at the election. We execute steps one to three separately for both parties. This allows us to account for different partisan mental models of the economy.

Inference Standard statistical inference is not feasible for synthetic beliefs. Since the synthetic belief is inspired by Abadie et al. (2015)'s synthetic-control method, we adapt some of their falsification exercises. We run in-time placebo studies to show both that the synthetic belief is a good predictor for perceptions in the absence of political change, and that structural breaks in perception formation can indeed be attributed to the 2016 presidential election. We also conduct "leave-one-out" exercises to assess the sensitivity of our results to the specific set of indicators selected by LASSO in the first step.

3.5.2 Results

Figure 3.5.3 compares stated economic perceptions to their synthetic counterparts. Republican perceptions (Panel (a)) increase discontinuously at the election threshold, and deviate markedly from the synthetic belief from then on. This is in line with our third prediction for motivated cognition. Compared to pre-election behavior, now-incumbent partisans react excessively to positive signals. Synthetic beliefs for Republicans do increase during the Republican presidency, suggesting some positive updating is indeed warranted given economic conditions. However, perceived improvements are by orders of magnitude larger than actual improvements. For Democrats (Panel (b)), the evidence is less conclusive. Synthetic beliefs and perceptions do not diverge in the immediate post-election period, although there is some downward divergence from Spring 2017 onwards. Recall that in our model, opposition partisans maximize their utility by holding as bad beliefs about the economy as possible given the signal structure. Since they cannot delude themselves completely, i.e., encode negative signals when signals are positive, their best option is to ignore positive information. In our case, the immediate post-election development of the economy was positive. The flat trajectory of Democrats' perceptions is in line with this prediction. The less volatile pre-election perceptions for Democrats suggest that different, longer-term indicators determine Democrats' perceptions compared to Republicans. We attribute the fact that the synthetic belief for Democrats only increases after a few months post-election to these indicators also only improving with a lag. We conclude that the results based on synthetic beliefs strongly support our hypothesis that partisan asymmetries in economic perceptions are due to motivated cognition.

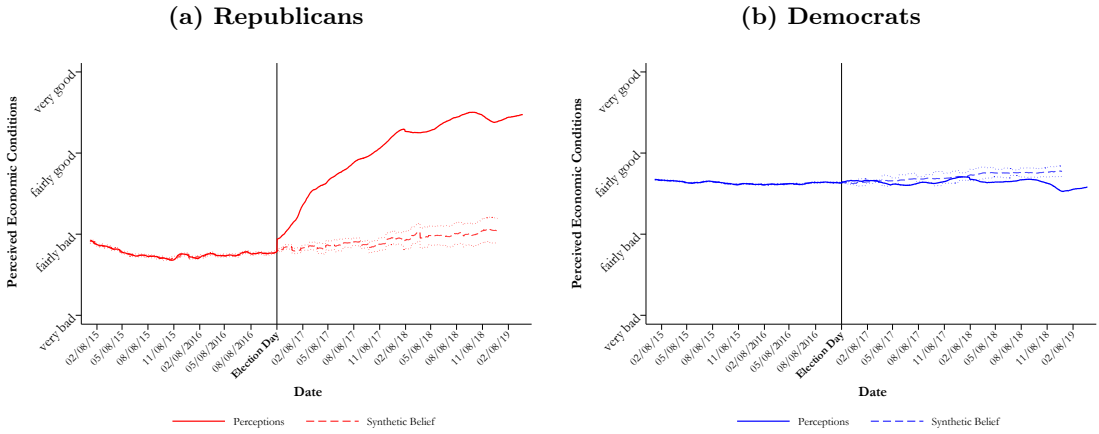
Discussion and Robustness Since standard statistical inference is not feasible for synthetic beliefs, we adapt the recommended sensitivity checks by Abadie et al. (2015) to our setting. We conduct several in–time placebo studies and compare synthetic beliefs of varying sparsity.

Placebo studies assess whether the divergence of synthetic beliefs and stated perceptions can indeed be attributed to the election, by re-estimating synthetic beliefs using an earlier cut-off date. This placebo election should not impact perceptions, such that synthetic beliefs closely mirror stated perceptions until the true election date.

Figure 3.5.4 shows results for a placebo election on May 3rd, 2016, when Donald J. Trump became presumptive nominee of the Republican party. For Republicans (Panel (a)), synthetic beliefs and perceptions begin to diverge at the November election, whereas for Democrats – comparable to Figure 3.5.3, Panel (b)) – synthetic beliefs only start to deviate from perceptions in spring 2017. Appendix Figures 3.B.4–3.B.6 show results for three further placebo elections. None lead to evidence against our findings.

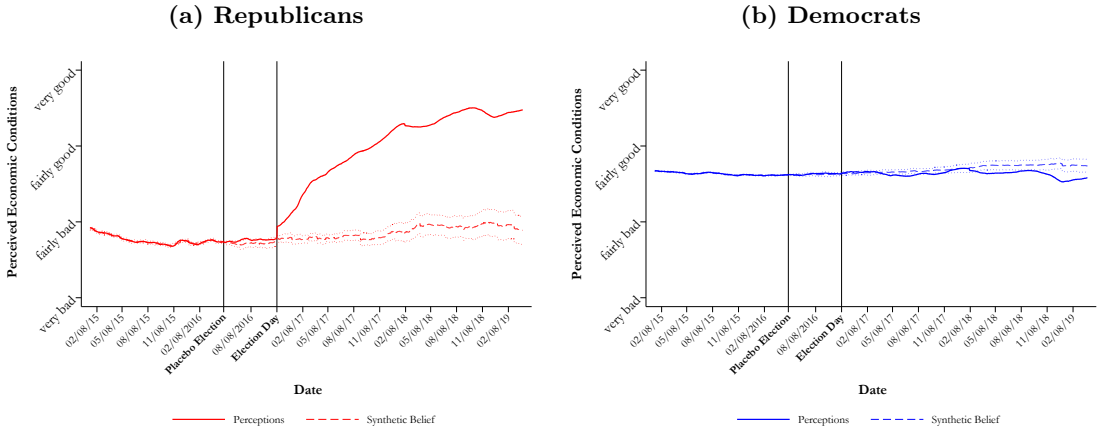
As a further robustness analysis, we replicate the analysis with "leave-one-out" synthetic beliefs.

Figure 3.5.3: Prediction 3: Excess Belief Movement at Power Shifts



Notes: The figure shows stated perceptions and synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) to test Prediction 3: Excess belief movement at power shifts. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Synthetic beliefs are calculated as described in Appendix 3.5. Dotted lines indicate the 95% confidence interval of auxiliary forecasting error. Since standard errors are not available for LASSO, forecasting error is error of post-LASSO OLS (c.f. Belloni and Chernozhukov (2013)).

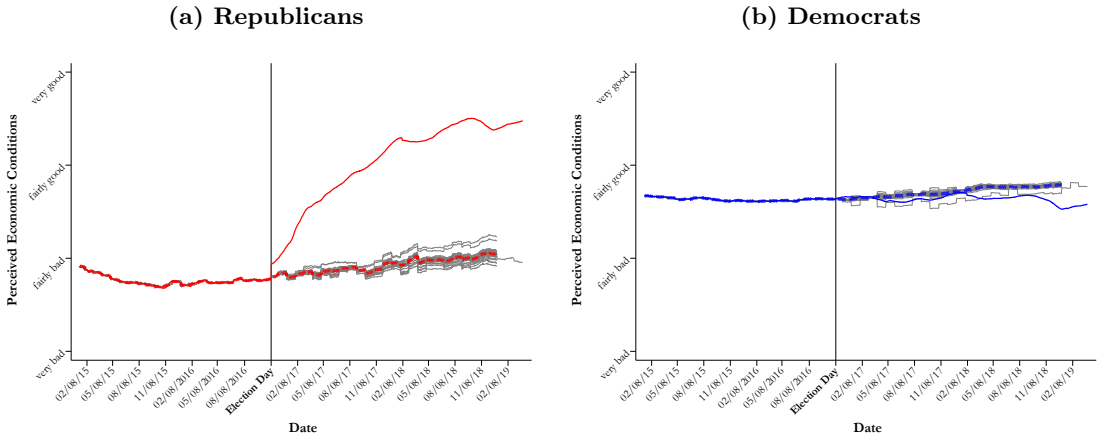
Figure 3.5.4: Placebo Election: Trump Primary Frontrunner



Notes: The figure shows stated perceptions and synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) for a placebo election on 05/03/2016 to test Prediction 3: Excess belief movement at power shifts. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Synthetic beliefs are calculated as described in Appendix 3.5. Dotted lines indicate the 95% confidence interval of auxiliary forecasting error. Since standard errors are not available for LASSO, forecasting error is error of post-LASSO OLS (c.f. Belloni and Chernozhukov (2013)).

These check whether results are driven by one particular indicator within the set of economic variables selected by LASSO. We consecutively drop one indicator from the set of best predictors, and then predict synthetic beliefs with the remaining $n-1$ indicators. The resulting synthetic beliefs serve as lower and upper bounds for the effects of the main synthetic beliefs.

Figure 3.5.5: Leave-One-Out Synthetic Beliefs



Notes: The figure shows stated perceptions, synthetic beliefs, and leave-one-out synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) to test Prediction 3: Excess belief movement at power shifts. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Synthetic beliefs are calculated as described in Appendix 3.5. Grey lines reflect Synthetic Beliefs predicted with $n-1$ indicators of the main synthetic belief (dashed line).

As shown in Figure 3.5.5, results from the main synthetic belief lie closely within the mass of leave-one-out synthetic beliefs. If at all, they suggest that our results are lower bounds for the extent of motivated cognition.

3.6 Discussion

In this section, we briefly discuss some alternative channels that could create the patterns we observe in the data.

Affect Prior et al. (2015) and Bullock et al. (2015) find that partisan disagreement about unemployment and inflation rates is reduced when survey participants are incentivized for correct responses. They conclude that partisan responses reflect affect – the emotion associated with the respective party or president – as well as genuinely held beliefs. However, the partisan gaps we observe remain significant, even if we adjust them to 40% of the initial partisan difference,

corresponding to the largest reduction in Prior et al. (2015) and Bullock et al. (2015). Thus, we are confident that our results are not entirely explained by affect.

Expectations Partisans might confound contemporaneous perceptions of the economy with their expectations for the future economy. Naturally, partisans have different beliefs over what policies will create economic prosperity. We do not directly control for expectations since they are a function of perceptions themselves. However, Civiqs elicits expectations in a separate question.[14] Appendix Figure 3.B.7 shows topline trends for partisans' expectations. The patterns are as expected, with Democrats being more optimistic about the future economy during a Democratic presidency, and Republicans more so during a Republican presidency. When we compare this to the patterns of perceptions in Figure 3.4.2, we are reassured that partisan differences in perceptions do not merely reflect differences in expectations, and that partisans are able to distinguish between the two questions. In particular, if expectations carried over to perceptions, we would expect to see a similar downward trend of Democrats' perceptions after the 2016 election as we see it for Democrats' expectations.

Biased Information Supply Larcinese et al. (2011) document agenda setting in U.S. newspapers: Partisan newspapers report economic news more favorably during incumbency than during opposition. Similarly, Bisgaard and Slothuus (2018) find that party-elite communication on economic issues has a strong impact on the beliefs of partisans. Partisan-biased information supply would generate similar patterns as partisan-biased information processing, that is, motivated cognition. Unfortunately, we cannot address this concern in our setting, as we do not observe media consumption of survey respondents. However, Gentzkow and Shapiro (2010) provide evidence that readers' demand for like-minded news drives media bias. In this vein, media bias and motivated beliefs go hand in hand: Both arise through preferences over beliefs, and reinforce each other.

3.7 Conclusion

In this paper, we investigate whether motivated cognition can explain partisan disagreement over facts. In a theoretical model, we derive three predictions as litmus tests for motivated cognition. These are all borne out in the dynamics of partisan perceptions around the 2016 U.S. presidential election: Republicans perceive economic conditions as better than Democrats during the Republican presidency, and the same holds true for Democrats during the Democratic presidency. Additionally, Republicans react excessively to increasing economic conditions after the Republican win in 2016.

[14] *National Economy: Direction.* The exact wording of the question is "Do you think the nation's economy is getting better or worse?" Possible answers are "getting better", "staying about the same", "getting worse", and "unsure".

Our study makes four contributions: First, we transfer the theoretical and experimental literature on motivated cognition to the field. Second, we make a methodological contribution by providing a novel method, the synthetic belief, for studying cognitive biases and inconsistencies in unstructured observational data. Third, we investigate the determinants of beliefs about the economy. Economic perceptions enter in a multitude of political economy models, most prominently in Fiorina (1981)'s theory of economic voting, in which perceived economic conditions determine an incumbent's reelection probability. Our results suggest that the opposite is also true: How we perceive the economy is influenced by who gets elected.

Finally, we add to the study of polarization by showing that observed patterns of polarized beliefs about facts are consistent with motivated cognition. In the future, it will be paramount to understand how motivated cognition, information provision, and politics in particular interact. To the extent that polarization reflects a fundamental human preference for congenial beliefs, it is unlikely that it can be mitigated by increased information provision or education.

Appendix to Chapter 3

- Appendix 3.A provides a list of economic time series used in Section 3.5.
- Appendix 3.B reports additional findings related to Sections 3.4 and 3.5.
- Appendix 3.C provides theoretical background and derivations for *Synthetic Beliefs* (Section 3.5).

3.A List of Used Economic Time Series and Sources

Table 3.A.1: List of Economic Data Series

Indicator	Source	Series Name in Source	Details	Frequency
		General Economic Conditions		
Leading Index for the United States	Philadelphia FED, from FRED	USSLIND	SA, percent	Monthly
ISM Manufacturing Index (PMI)	Quandl	PMI	SA, %	Monthly
ISM Nonmanufacturing Index (NMI)	Quandl	NMI	SA, %	Monthly
Industrial Production	BOG, from FRED	IPMAN	SA, index base 2012, NAICS manufacturing	Monthly
GDP	BEA, Table 8.1.5	-	Not SA	Quarterly
Gross Domestic Income	BEA, Table 8.2	-	Not SA	Quarterly
Savings Rate	BEA, Table 5.1	-	SA, % of GNI	Quartely
Net Saving	BEA, Table 5.1	-	SA	Quartely
Gross Saving	BEA, Table 5.1	-	SA	Quartely
Consumption	BEA, Table 8.1.5	-	SA	Quarterly
Domestic Investment	BEA Table 8.2	-	Not SA, gross	Quarterly
		Households and Housing		
Personal Income	BEA, Table 2.6	-	SA	Monthly
Personal Saving	BEA, Table 2.6	-	SA, % of disposable income	Monthly
Housing Starts	Census Bureau & HUD, from FRED	HOUST	SA	Monthly
New One-Family Homes	Census Bureau & HUD, from FRED	HSN1F	SA, sold units	Monthly
Auto Sales	BEA, from FRED	ALTSALES	SA, incl. light trucks	Monthly
		Labor Market		
Coincident Economic Activity Index	Philadelphia FED, from FRED	USPHCI	SA, base 2012	Monthly
Population	BEA, from FRED	POPTHM	-	Monthly
Working Age Population	OECD, from FRED	LFWA64TTUSM647S	SA, Ages 15-64	Monthly
Labor Force	BLS	LNS11000000	SA, Civilian, Ages 16+	Monthly
Employment Rate	OECD, from FRED	LREM64TTUSM156S	SA, Ages 15-64	Monthly
Employment Level	BLS	LNS12000000	SA, Ages 16+	Monthly
Private Employees	BLS	CES0500000001	SA	Monthly
Non-Farm Employees	BLS	CES0000000001	SA	Monthly
Unemployment Rate	BLS	LNS1400000	SA, Ages 16+	Monthly
Unemployment Level	BLS	LNS1300000	SA, Ages 16+	Monthly
Unemployment Level	BLS	LNU04000000	Not SA, Ages 16+	Monthly
Hourly Earnings, Production	BLS	CES0500000008	SA, all nonsupervisory	Monthly
Hourly Earnings	BLS	CES0500000003	SA, private employees	Monthly

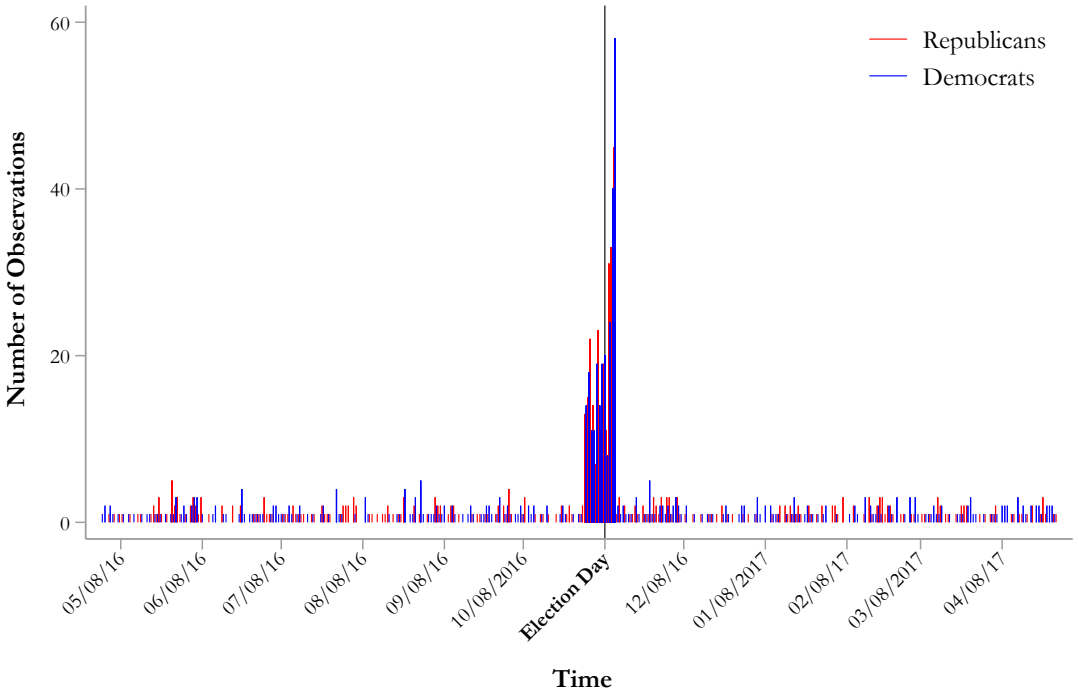
Compensation	BLS	PRS85006112	Change in real hourly rate	Quarterly
Weekly Hours, Production	BLS	CES0500000007	SA, all nonsupervisory	Monthly
Weekly Hours	BLS	CES0500000002	SA, private employees	Monthly
		Firms		
Profits	BEA, Table 6.16 D	-	SA	Quarterly
Domestic Profits	BEA Table 6.16 D	-	SA	Quarterly
Industrial Production Index	BOG, from FRED	INDPRO	SA, base 2012	Monthly
Manufacturing Profits	BEA, Table 6.16 D	-	SA	Quarterly
Labor Productivity	BLS	PRS85006092	Non-Farm, % Change	Quarterly
Labor Productivity	BLS	PRS30006092	Manufacturing, % Change	Quarterly
Multifactor Productivity	BLS	MPU4910012	Index, base 2012	Annual
Labor Cost	BLS	PRS85006112	Non-Farm, % Change	Quarterly
Employment Cost Index (ECI)	BLS	CIU1010000000000A	Not SA, 12 month change	Quarterly
Domestic Business Investment	BEA Table 8.2	-	Not SA, gross	Quarterly
Business Investment, Structures	BEA, Table 8.1.5	-	Not SA	Quarterly
Business Investment, Equipment	BEA, Table 8.1.5	-	Not SA	Quarterly
Business Investment, Intellectual Property	BEA, Table 8.1.5	-	Not SA	Quarterly
Business Investment, Residential	BEA, Table 8.1.5	-	Not SA	Quarterly
Retail Sales	Census Bureau, from FRED	MRTSMPCSM44X72USS	SA, % Change, total incl. food	Quarterly
Retail Sales excl. Autos	Census Bureau, from FRED	MRTSMPCSM44Y72USS	SA, % Change, total excl. vehicles	Quarterly
New Orders, Manufacturing	Census Bureau, from FRED	NEWORDER	SA, excl. aircraft & defense goods	Monthly
New Orders of Durables	Census Bureau, from FRED	DGORDER	SA	Monthly
Inventories to Sales Ratio, Retail	Census Bureau, from FRED	RETAILIRNSA	Not SA, ratio	Monthly
Inventories to Sales Ratio, Manufacturing	Census Bureau, from FRED	MNFCTRIRNSA	Not SA, ratio	Monthly
Inventories to Sales Ratio, Wholesalers	Census Bureau, from FRED	WHLSLRIRNSA	Not SA, ratio	Monthly
Inventories to Sales Ratio, Total	Census Bureau, from FRED	TOTBUSIRNSA	Not SA, ratio	Monthly
Total Business Inventory	Census Bureau, from FRED	TOTBUSMPCIMSA	SA, % Change	Monthly
Capacity Utilization, Manufacturing	BOG, from FRED	MCUMFN	SA, NAICS Manufacturing	Monthly
Capacity Utilization	BOG, from FRED	TCU	SA, all industry, percent	Monthly
		Stock Market		
10-Y HQM Corporate Bond Rate	USDT, from FRED	HQMCB10YR	Not SA, calculated by USDT	Monthly
S&P 500	S&P, from FRED	SP500	Not SA	Daily
Dow Jones Composite Average	S&P, from FRED	DJCA	Not SA	Daily
Dow Jones Industrial Average	S&P, from FRED	DJIA	Not SA	Daily

		Trade		
Trade Balance, Goods and Services	BEA and Census Bureau, from FRED	BOPGSTB	SA	Monthly
Exports	BEA, Table 8.1.5	-	Not SA	Quarterly
Imports	BEA, Table 8.1.5	-	Not SA	Quarterly
Trade Balance, Goods	BEA & Census Bureau, from FRED	BOPGTB	SA	Monthly
Import Price Index	BLS	EIUIR	Not SA, base 2000	Monthly
Export Price Index	BLS	EIUIQ	Not SA, base 2000	Monthly
Trade-Weighted US Dollar Index	BOG, from FRED	DTWEXM	Not SA, goods, base 1973	Daily
Exchange Rate: China	BOG, from FRED	DEXCHUS	Not SA, CNY/$	Daily
		Government		
Federal Government Expenditure	BEA, Table 8.1.5	-	Not SA	Quarterly
Total Federal Debt	USDT, from FRED	GFDEBTN	Not SA	Quarterly
New Government Debt	BEA, Table 3.1	-	SA, net lending borrowing	Quarterly
Government Debt Quota	FRED	GFDEGDQ188S	SA, federal, % of GDP	Quarterly
State & Local Government Expenditure	BEA, Table 8.1.5	-	Not SA	Quarterly
Government Revenue	BEA, Table 3.1	-	SA	Quarterly
Government Consumption	BEA, Table 3.9.5	-	SA	Quarterly
Government Investment	BEA, Table 3.9.5	-	SA	Quarterly
		Interest Rates		
Effective Federal Funds Rate	BOG, from FRED	FEDFUNDS	Not SA	Monthly
3-M Treasury Bill	BOG, from FRED	DTB3	Not SA, secondary market rate	Daily
10-Y Treasury Bill	BOG, from FRED	DGS10	Not SA, constant maturity rate	Daily
LIBOR Rate	IBA, from FRED	USD3MTD156N	Not SA, rate	Daily
TED Spread	FRED	TEDRATE	Not SA, %	Daily
		Prices		
Inflation as Δ CPI	BLS	CUUR0000SA0	base 1982-1984, urban consumers, own calculation	Monthly
CPI excl. food & energy	BLS	CUUR0000SA0L1E	Not SA, base 1982-1984	Monthly
PPI commodity	BLS	WPSFD4	SA, base 11/2009	Monthly
PPI finished goods	BLS	WPUFD49207	Not SA, base 1982	Monthly
Inflation Expectation	FRED	T5YIFR	Not SA, 5 years forward	Daily
PCE	BEA, Table 2.8.7	-	SA	Monthly
PCE excl. food & energy	BEA, Table 2.3.7	-	SA	Monthly
GDP Deflator	BEA, Table 8.1.4	-	Not SA, base 2012	Quarterly
Spot Oil Price	FRED	WTISPLC	Not SA, WTI Crude, $/barrel	Monthly

Notes: Abbreviations: BEA (U.S. Bureau of Economic Analysis), BLS (U.S. Bureau of Labor Statistics), BOG (Board of Governors of the Federal Reserve System), HUD (U.S. Department of Housing and Urban Development), IBA (ICE Benchmark Administration Limited), USDT (U.S. Department of the Treasury), S&P (S&P Dow Jones Indices LLC), SA (Seasonally adjusted), all variables levels of real unit counts unless otherwise specified.

3.B Additional Figures and Tables

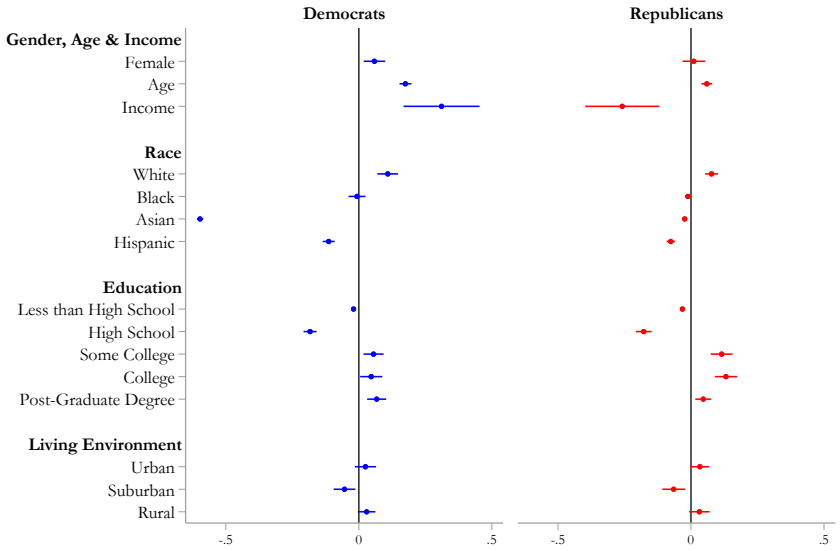
Figure 3.B.1: Distribution of Observations in Survey Sample



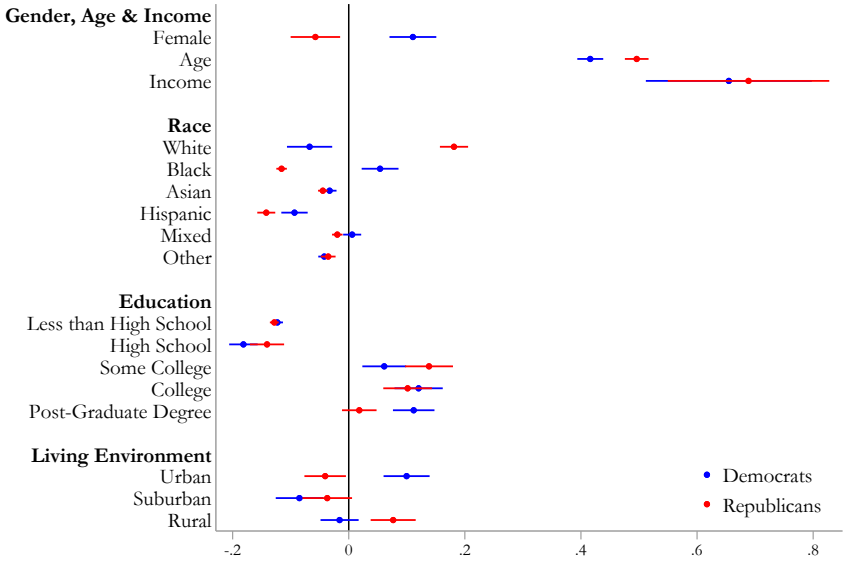
Notes: The figure displays the distribution of respondents in the individual-level survey sample by day, data covers 05/01/2016–04/29/2017.

Figure 3.B.2: Representativeness of Survey Sample

(a) Relative to Partisan Population

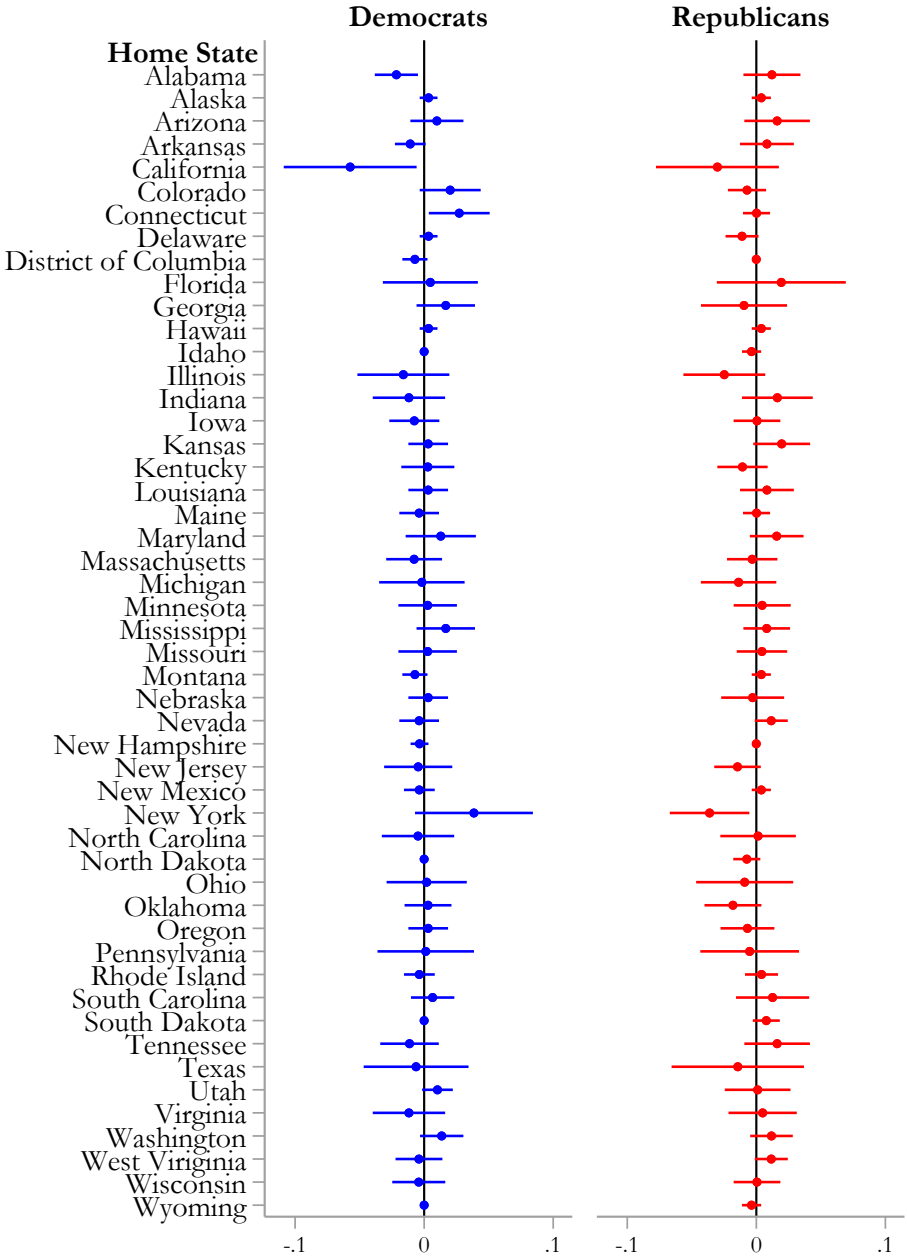


(b) Relative to U.S. Population



Notes: The figure displays differences between sample and U.S. population characteristics. Female is an indicator that is equal to one if the respondent self-identifies as female (1) or male (0). Third gender/non-binary gender is coded as missing due to few observations. Age is a respondent's age, measured in years. Income is reported in units of $25,000. White/Black/Asian/Hispanic/Other are indicator variables that are equal to one if the respondent identifies with a given ethnicity. Ethnicities with fewer than 20 observations are included in "Other". Education variables are indicators that are equal to one if the respondent falls into a given educational category. Urban is an indicator that is equal to one if the respondent lives in an urban area, suburban is an indicator that is equal to one if the respondent lives in a suburban area, and rural is an indicator that is equal to one if the respondent lives in a rural area. Lines indicate 95% confidence intervals. Population and party baseline taken from Pew Research Center (2016, 2018); U.S. Census Bureau (2018) and own calculations.

Figure 3.B.3: Pre-and Post Election Characteristics of Survey Sample: States



Notes: Difference in pre- and post-election home state of partisans. Lines indicate 95% confidence intervals.

Table 3.B.2: Summary Statistics for Individual-Sample Survey Data

	Democrats	Republicans	Total
National Economy: Condition	1.91	0.84	1.39
	(0.71)	(0.75)	(0.91)
	551	521	1072
Female	0.62	0.45	0.54
	(0.49)	(0.50)	(0.50)
Observations	561	526	1087
Age	59.14	63.63	61.31
	(13.39)	(12.92)	(13.35)
Observations	555	519	1074
Income	3.21	3.25	3.23
	(1.58)	(1.50)	(1.54)
Observations	470	450	920
Race			
White	0.66	0.91	0.78
	(0.47)	(0.28)	(0.41)
Black	0.18	0.01	0.10
	(0.39)	(0.11)	(0.30)
Asian	0.02	0.01	0.02
	(0.14)	(0.10)	(0.12)
Hispanic	0.08	0.03	0.06
	(0.28)	(0.18)	(0.24)
Mixed	0.04	0.01	0.02
	(0.19)	(0.11)	(0.15)
Other	0.02	0.02	0.02
	(0.12)	(0.15)	(0.14)
Observations	569	530	1099
Education			
< High School	0.01	0.01	0.01
	(0.11)	(0.09)	(0.10)
High-School Graduate	0.09	0.13	0.11
	(0.29)	(0.34)	(0.31)
< College	0.27	0.35	0.31
	(0.44)	(0.48)	(0.46)
College Graduate	0.39	0.38	0.39
	(0.49)	(0.48)	(0.49)
College +	0.23	0.14	0.18
	(0.42)	(0.34)	(0.39)
Observations	535	514	1049
Living Environment			
Urban	0.37	0.23	0.30
	(0.48)	(0.42)	(0.46)
Suburban	0.44	0.48	0.46
	(0.50)	(0.50)	(0.50)
Rural	0.20	0.29	0.24
	(0.40)	(0.45)	(0.43)
Observations	570	530	1100

Notes: The table reports summary statistics for the survey sample. Female is an indicator that is equal to one if the respondent self-identifies as female (1) or male (0). Third gender/non-binary gender is coded as missing due to few observations. Age is a respondent's age, measured in years. Income is reported in units of $25,000. White/Black/Asian/Hispanic/Other are indicator variables that are equal to one if the respondent identifies with a given ethnicity. Ethnicities with fewer than 20 observations are included in "Other". Education variables are indicators that are equal to one if the respondent falls into a given educational category. Urban is an indicator that is equal to one if the respondent lives in an urban area, suburban is an indicator that is equal to one if the respondent lives in a suburban area, and rural is an indicator that is equal to one if the respondent lives in a rural area. Standard deviations in parentheses.

Table 3.B.3: Prediction 1: Partisan Disagreement, All Controls

Dependent Variable	Perceptions	
	(1)	(2)
Republican Presidency × Republican	0.3419***	0.3628***
	(0.0934)	(0.0952)
Republican Presidency	-0.1558*	-0.1801**
	(0.0864)	(0.0905)
Republican	-1.2148***	-1.2401***
	(0.0694)	(0.0739)
Dow Jones CA	0.0004***	0.0005***
	(0.0001)	(0.0001)
Female	-0.0475	-0.0542
	(0.0469)	(0.0488)
Age	-0.0007	-0.0018
	(0.0019)	(0.0019)
Income	0.0352**	0.0412**
	(0.0164)	(0.0171)
Asian	0.6658***	0.8446***
	(0.2338)	(0.2454)
Black	0.0760	0.0707
	(0.1889)	(0.2028)
Hispanic	0.0495	0.0225
	(0.1994)	(0.2107)
Mixed	-0.1321	-0.0929
	(0.2178)	(0.2320)
White	-0.0802	-0.0828
	(0.1732)	(0.1844)
Post-Graduate Degree	-0.0737	-0.0541
	(0.2986)	(0.3375)
College	-0.0390	-0.0003
	(0.2961)	(0.3324)
Some College	-0.0938	-0.0261
	(0.2968)	(0.3331)
High School	-0.1913	-0.1264
	(0.3016)	(0.3381)
Urban	-0.0432	-0.0461
	(0.0571)	(0.0613)
Rural	-0.0908	-0.0990
	(0.0581)	(0.0609)
Constant	-0.7832	-0.7858
	(0.8007)	(0.8649)
State FEs		Yes
Adjusted R^2	0.42	0.42
Observations	874	874

Notes: The table reports results for Equation (3.4.1) testing Prediction 1: Partisan Disagreement irrespective of economic conditions. The dependent variable are perceptions measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?", responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing). Female contains whether the respondent self-identifies as female (1) or male (0). Third gender/non-binary gender coded as missing due to few observations. Ethnicities with fewer than 20 observations are included in "Other". Income is reported in units of $25,000. Age is measured in years. Suburban, "Other" ethnicity, and less than high school education are omitted categories. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3.B.4: Prediction 2: Selective Relevance of Economic Information, by Partisanship, All Controls

(a) Republicans

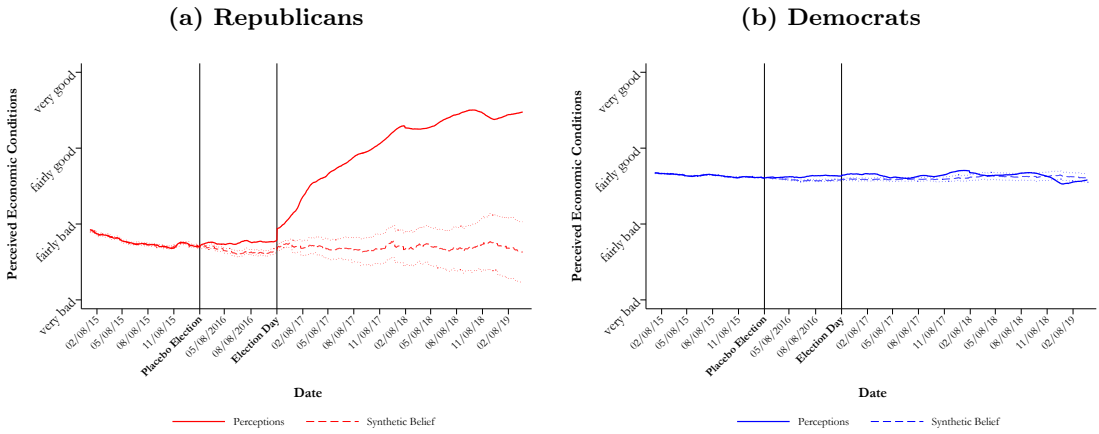
Dependent Variable	Perceptions	
	(1)	(2)
Republican Presidency × Dow Jones CA	0.0012**	0.0013**
	(0.0006)	(0.0006)
Republican Presidency	-7.7885**	-8.2444**
	(3.6148)	(3.8514)
Dow Jones CA	-0.0000	-0.0001
	(0.0005)	(0.0006)
Female	-0.0777	-0.0717
	(0.0685)	(0.0721)
Age	-0.0049	-0.0058*
	(0.0030)	(0.0031)
Income	0.0133	0.0127
	(0.0250)	(0.0265)
Asian	1.0355***	1.1079***
	(0.2864)	(0.2878)
Black	0.3509	0.3136
	(0.4409)	(0.5047)
Hispanic	0.5024*	0.4729
	(0.2751)	(0.3257)
Mixed	0.2546	0.2207
	(0.3232)	(0.3254)
White	0.3124*	0.2608
	(0.1803)	(0.1991)
Post-Graduate Degree	0.2915	0.3728
	(0.4563)	(0.4918)
College	0.2166	0.3124
	(0.4485)	(0.4804)
Some College	0.1187	0.1955
	(0.4474)	(0.4801)
High School	0.0236	0.1623
	(0.4516)	(0.4843)
Urban	-0.0784	-0.1001
	(0.0915)	(0.0991)
Rural	-0.0452	-0.0716
	(0.0743)	(0.0856)
Constant	0.5985	0.9881
	(3.4715)	(3.6426)
State FEs		Yes
Adjusted R^2	0.16	0.14
Observations	432	432

(b) Democrats

Dependent Variable	Condition	
	(1)	(2)
Republican Presidency × Dow Jones CA	0.0004	0.0004
	(0.0005)	(0.0005)
Republican Presidency	-2.3151	-2.5718
	(3.1590)	(3.3171)
Dow Jones CA	-0.0004	-0.0005
	(0.0005)	(0.0005)
Female	-0.0496	-0.0705
	(0.0628)	(0.0673)
Age	0.0024	0.0021
	(0.0024)	(0.0025)
Income	0.0484**	0.0714***
	(0.0218)	(0.0238)
Asian	0.1823	0.3516
	(0.2833)	(0.3200)
Black	-0.4073*	-0.3848
	(0.2327)	(0.2674)
Hispanic	-0.4660*	-0.3810
	(0.2442)	(0.2739)
Mixed	-0.6027**	-0.4260
	(0.2717)	(0.2965)
White	-0.5748**	-0.5398**
	(0.2227)	(0.2544)
Post-Graduate Degree	-0.3279	-0.2116
	(0.3103)	(0.3617)
College	-0.2311	-0.0866
	(0.3073)	(0.3557)
Some College	-0.2246	-0.0321
	(0.3094)	(0.3592)
High School	-0.3120	-0.1416
	(0.3257)	(0.3734)
Urban	-0.0074	0.0355
	(0.0707)	(0.0841)
Rural	-0.1128	-0.0839
	(0.0901)	(0.0976)
Constant	5.2801*	5.5570*
	(2.9796)	(3.2206)
State FEs		Yes
Adjusted R^2	0.03	0.04
Observations	442	442

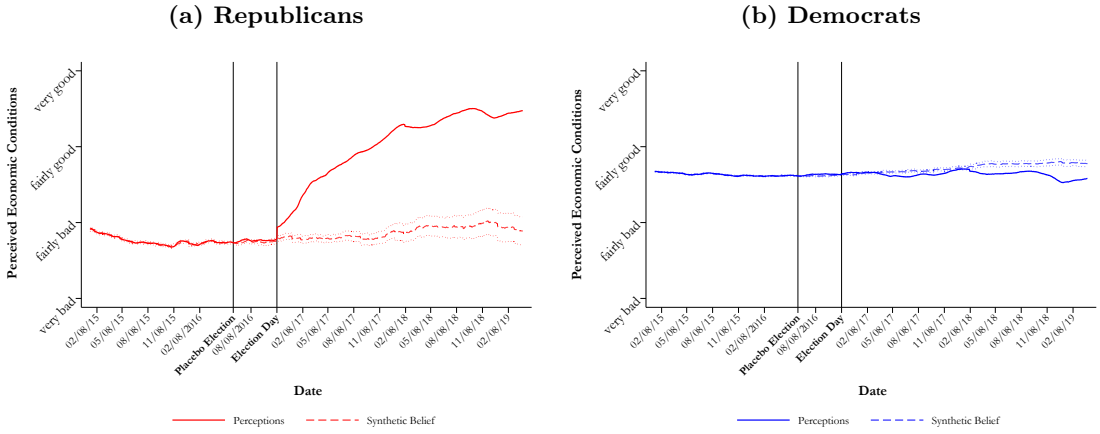
Notes: The table reports results for Equation (3.4.2), testing Prediction 2: Partisan Disagreement sustained by selective relevance of economic information. The dependent variable are perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?", responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing). Panel (a) displays results for Republicans, and Panel (b) displays results for Democrats. Female contains whether the respondent self-identifies as female (1) or male (0). Third gender/non-binary gender coded as missing due to few observations. Ethnicities with fewer than 20 observations are included in "Other". Income is reported in units of $25,000. Age is measured in years. Suburban, "Other" ethnicity, and less than high school education are omitted categories. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Figure 3.B.4: Placebo Election: February 2016



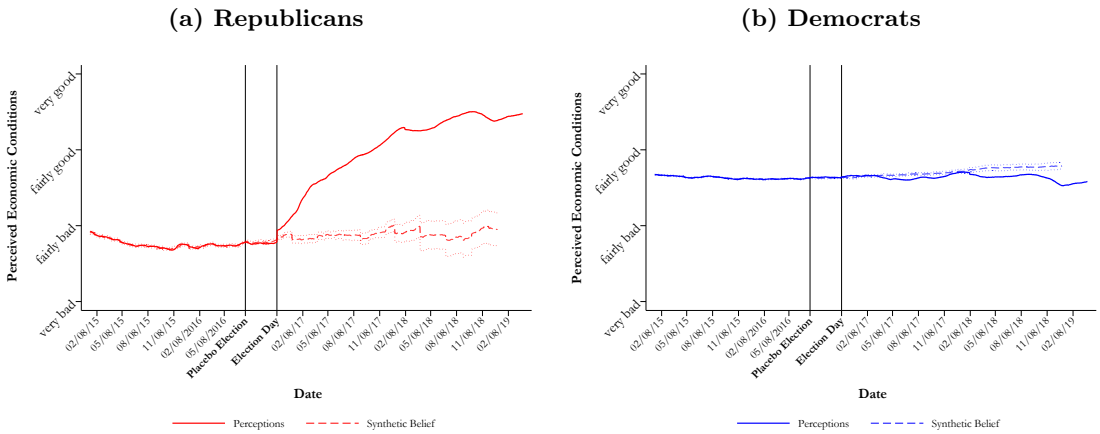
Notes: The figure shows stated perceptions and synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) for a placebo election on 02/08/2016 to test Prediction 3: Excess belief movement at power shifts. Comparison of synthetic beliefs and stated perceptions. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Dotted lines indicate the 95% confidence interval of auxiliary forecasting error. Since standard errors are not available for LASSO, forecasting error is error of post-LASSO OLS (c.f. Belloni and Chernozhukov (2013)).

Figure 3.B.5: Placebo Election: Clinton Primary Frontrunner



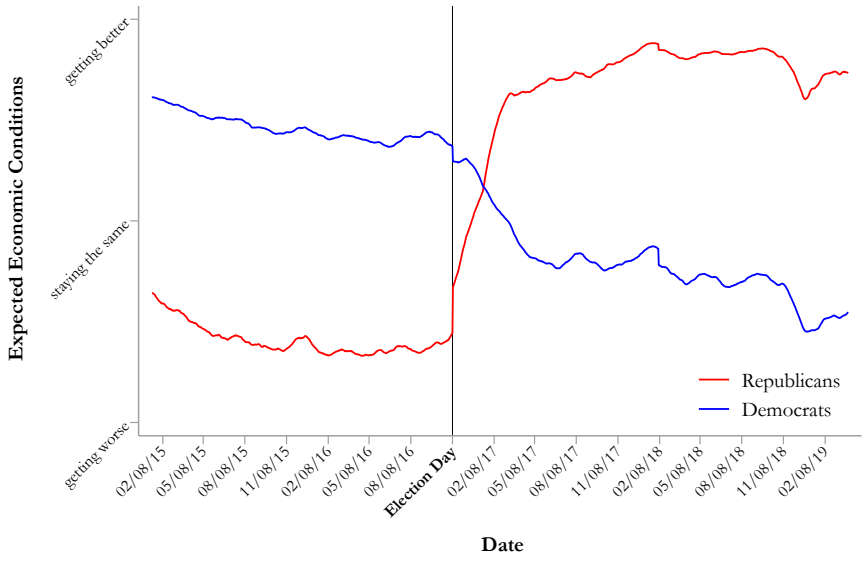
Notes: The figure shows stated perceptions and synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) for a placebo election on 06/06/2016 to test Prediction 3: Excess belief movement at power shifts. Comparison of synthetic beliefs and stated perceptions. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Dotted lines indicate the 95% confidence interval of auxiliary forecasting error. Since standard errors are not available for LASSO, forecasting error is error of post-LASSO OLS (c.f. Belloni and Chernozhukov (2013)).

Figure 3.B.6: Placebo Election: RNC/DNC



Notes: The figure shows stated perceptions and synthetic beliefs for Republicans (Panel (a)) and Democrats (Panel (b)) for a placebo election on 07/19/2016 to test Prediction 3: Excess belief movement at power shifts. Comparison of synthetic beliefs and stated perceptions. Perceptions are measured as survey responses to *National Economy: Current Condition*: "How would you rate the condition of the national economy right now?" partisan aggregate time series from Civiqs.com, responses re-coded as: "very good" (3), "fairly good" (2), "fairly bad" (1), "very bad" (0), "unsure" (missing), and weighted by percentage giving each answer. Dotted lines indicate the 95% confidence interval of auxiliary forecasting error. Since standard errors are not available for LASSO, forecasting error is error of post-LASSO OLS (c.f. Belloni and Chernozhukov (2013)).

Figure 3.B.7: Aggregate Trends in Economic Expectations



Notes: The figure shows survey responses to *National Economy: Direction*: "Do you think the nation's economy is getting better or worse?", "getting better" (+1), "staying about the same"(0), "getting worse" (−1), and "unsure" (missing), partisan aggregate time series from Civiqs.com, responses re-coded and weighted by percentage giving each answer, data covers 01/15/2015 to 03/31/2019.

3.C Theoretical Model for Synthetic Belief

There are $N = N^I + N^O$ agents, i, belonging to party $\rho \in (I, O)$, where I is the incumbent party and O the opposition, who each form perception, e^ρ_{it}, at time t. e^ρ_{it} is a function of priors p^ρ_{0it}, a vector of k signals, $\mathbf{s_{it}}$, about K economic indicators, k_t, at time t, and a $k+1$ vector of value weights, $\mathbf{v}^\rho$, that determine how much weight an individual puts on any indicator – including their prior – when assessing the condition of the economy. We do not allow for individual differences in value weights. Instead, we assume that value weights are how partisanship enters non-motivated perceptions: There is partisan disagreement about what constitutes good economic conditions, but the same values are shared by all members of a party (*Shared-Values Assumption*).

$$e^\rho_{it} = f(\mathbf{v}^\rho, p^\rho_{0it}, \mathbf{s_{it}})$$

If function $f(\mathbf{v}^\rho, p^\rho_{0it}, \mathbf{s_{it}})$, value weights, priors and individual signals were known, we could predict non-motivated perceptions from signals. Our goal is therefore to calibrate $f(\mathbf{v}^\rho, p^\rho_{0it}, \mathbf{s_{it}})$. For simplicity, we assume that f can be approximated by a linear function:

$$e^\rho_{it} = f(\mathbf{v}^\rho, p^\rho_{0it}, \mathbf{s_{it}}) = v^\rho_{p0} p^\rho_{0it} + v^\rho_1 s_{1it} + v^\rho_2 s_{2it} + ... + v^\rho_k s_{kit}.$$

Additionally, we assume that individual priors, p^ρ_{0it}, are distributed iid. around a partisan mean $\bar{\rho}^r$. s_{kit} is the realization of signal s about economic indicator k at time t for individual i. We assume that signals, s_{kit}, are subject to random noise and thus iid. distributed with k_t as the distribution mean, such that aggregate assessments will reflect the true value of k_t (*Collective Intelligence Assumption*). This is an assumption commonly made by political scientists and on research on public opinion and is derived from the Condorcet jury theorem. For a discussion, see Duch et al. (2000).

Since, we do not observe individual information sets, we cannot estimate f from individual data. However, we can aggregate over all N^r supporters of a party:

$$\begin{aligned} E^\rho_t &= \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} e^\rho_{it} \\ &= \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} v^\rho_{p0} p^\rho_{0it} + \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} v^\rho_1 s_{1it} + \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} v^\rho_2 s_{2it} + ... + \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} v^\rho_k s_{kit} \\ &= v^\rho_{p0} \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} p^\rho_{0it} + v^\rho_1 \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} s_{1it} + v^\rho_2 \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} s_{2it} + ... + v^\rho_k \frac{1}{N^\rho} \sum_{i=1}^{N^\rho} s_{kit}. \end{aligned}$$

According to the Weak Law of Large Numbers, for sufficiently large N^ρ: $\frac{1}{N^\rho}\sum_{i=1}^{N^\rho} s_{kit} = k_t$ and $\frac{1}{N^\rho}\sum_{i=1}^{N^\rho} p^\rho_{0it} = p^\rho_0$:

$$E^\rho_t = v^\rho_{p0} p^\rho_0 + v^\rho_1 k_{1t} + v^\rho_2 k_{2t} + ... + v^\rho_k k_{kt}.$$

We can now calibrate value weights using stated perceptions and economic indicators. In the absence of changes in f, predicted perceptions $\hat{E^\rho_t}$ will converge to observed perceptions for sufficiently large N. We thus interpret differences between predicted perceptions – synthetic beliefs – and observed perceptions, as structural changes in the way perceptions are formed.

Bibliography

Abadie, A., A. Diamond, and J. Hainmueller (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333.

Abramitzky, R., L. Boustan, K. Eriksson, J. Feigenbaum, and S. Pérez (2021a): "Automated Linking of Historical Data," *Journal of Economic Literature*, 59, 865–918.

Abramitzky, R., L. Boustan, E. Jacome, and S. Perez (2021b): "Intergenerational Mobility of Immigrants in the United States over Two Centuries," *American Economic Review*, 111, 580–608.

Abramitzky, R., L. P. Boustan, and K. Eriksson (2012): "Europe's Tired, Poor, Huddled Masses: Self-selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102, 1832–1856.

Abramitzky, R., L. Greska, S. Pérez, J. Price, C. Schwarz, and F. Waldinger (2024a): "Climbing the Ivory Tower: How Socio-Economic Background Shapes Academia," *National Bureau of Economic Research Working Paper*.

Abramitzky, R., J. K. Kowalski, S. Pérez, and J. Price (2024b): "The GI Bill, Standardized Testing, and Socioeconomic Origins of the US Educational Elite Over a Century," *National Bureau of Economic Research Working Paper*.

Adhvaryu, A., N. Kala, and A. Nyshadham (2023): "Returns to On-the-Job Soft Skills Training," *Journal of Political Economy*, 131, 2165–2208.

Agarwal, R. and P. Gaule (2020): "Invisible Geniuses: Could the Knowledge Frontier Advance Faster?" *American Economic Review: Insights*, 2, 409–24.

Aghion, P., U. Akcigit, A. Hyytinen, and O. Toivanen (2018): "On the Returns to Invention within Firms: Evidence from Finland," *AEA Papers and Proceedings*, 108, 208–212.

——— (2023): "Parental Education and Invention: the Finnish Enigma," *National Bureau of Economic Research Working Paper*.

Agrawal, A., A. Goldfarb, and F. Teodoridis (2016): "Understanding the Changing Structure of Scientific Inquiry," *American Economic Journal: Applied Economics*, 8, 100–128.

Ahmadpoor, M. and B. F. Jones (2019): "Decoding team and individual impact in science and invention," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 13885–13890.

Airoldi, A. and P. Moser (2024): "Inequality in Science: Who Becomes a Star?" *mimeo NYU*.

Akcigit, U., J. Grigsby, and T. Nicholas (2017): "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age," *National Bureau of Economic Research Working Paper*.

Akerlof, G. A. (1989): "The Economics of Illusion," *Economics and Politics*, 1, 1–15.

Akerlof, G. A. and R. E. Kranton (2010): *Identity economics: How our identities shape our work, wages, and well–being*, Princeton: Princeton University Press.

Albrecht, J., T. Qin, and M. Yates (2024): "NeurIPS 2024 Call for Competitions," https://blog.neurips.cc/2024/03/03/neurips-2024-call-for-competitions/, last retrieved 2025/02/27.

Alesina, A., A. Miano, and S. Stantcheva (2020): "The Polarization of Reality," *AEA Papers and Proceedings*, 110, 324–28.

Allocca, A. (2024): "An Empirical Study of Team Formation and Performance," *mimeo University of Munich*.

Association for Computing Machinery (2025): "KDD Cup Archives," https://kdd.org/kdd-cup, last retrieved 2025/02/25.

Atomic Heritage Foundation (2022): "Raymond E. Zirkle," https://ahf.nuclearmuseum.org/ahf/profile/raymond-e-zirkle/, last retrieved 2024/12/09.

Azoulay, P., J. S. Graff Zivin, and J. Wang (2010): "Superstar Extinction," *The Quarterly Journal of Economics*, 125, 549–589.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017): "Gender differences in accepting and receiving requests for tasks with low promotability," *American Economic Review*, 107, 714–47.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017): "Does the gender composition of scientific committees matter?" *American Economic Review*, 107, 1207–38.

Bartels, B. L., J. M. Box-Steffensmeier, C. D. Smidt, and R. M. Smith (2011): "The dynamic properties of individual-level party identification in the United States," *Electoral Studies*, 30, 210–222.

Bartels, L. M. (2002): "Beyond the Running Tally: Partisan Bias in Political Perceptions," *Political Behavior*, 24, 117–150.

Beadle, G. W. (1974): "Recollections," *Annual Review of Biochemistry*, 43, 1–14.

Becker, G. S. and K. M. Murphy (1992): "The Division of Labor, Coordination Costs, and Knowledge," *The Quarterly Journal of Economics*, 107, 1137–1160.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019): "Who becomes an Inventor in America? The Importance of Exposure to Innovation," *The Quarterly Journal of Economics*, 134, 647–713.

Belloni, A. and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Bénabou, R. and J. Tirole (2002): "Self–Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 117, 871–915.

——— (2006): "Belief in a Just World and Redistributive Politics," *The Quarterly Journal of Economics*, 121, 699–746.

——— (2011): "Identity, Morals, and Taboos: Beliefs as Assets," *The Quarterly Journal of Economics*, 126, 805–855.

——— (2016): "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 30, 141–164.

Bennet, J. and S. Lanning (2007): "The Netflix Prize," *Proceedings of the KDD Cup and Workshop 2007*.

Bisgaard, M. and R. Slothuus (2018): "Partisan Elites as Culprits? How Party Cues Shape Partisan Perceptual Gaps," *American Journal of Political Science*, 62, 456–469.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020): "Are Ideas Getting Harder to Find?" *American Economic Review*, 110, 1104–44.

Bolton, P. and M. Dewatripont (1994): "The Firm as a Communication Network," *The Quarterly Journal of Economics*, 109, 809–839.

Bonhomme, S. (2022): "Teams: Heterogeneity, Sorting and Complementarity," *mimeo University of Chicago*.

Boudreau, K. J., N. Lacetera, and K. R. Lakhani (2011): "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management Science*, 57, 843–863.

Boudreau, K. J., K. R. Lakhani, and M. Menietti (2016): "Performance responses to competition across skill levels in rank-order tournaments: field evidence and implications for tournament design," *The RAND Journal of Economics*, 47, 140–165.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017): "Greater Internet use is not associated with faster growth in political polarization among US demographic groups," *Proceedings of the National Academy of Sciences of the United States of America*, 114, 10612–10617.

——— (2024): "Cross-Country Trends in Affective Polarization," *The Review of Economics and Statistics*, 106, 557–565.

Buckles, K., A. Haws, J. Price, and H. E. Wilbert (2023): "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project," *National Bureau of Economic Research Working Paper*.

Bullock, J. G., A. S. Gerber, S. J. Hill, and G. A. Huber (2015): "Partisan Bias in Factual Beliefs about Politics," *Quarterly Journal of Political Science*, 10, 519–578.

Bulten, W., K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, C. Hulsbergen-van de Kaa, J. van der Laak, M. B. Amin, A. J. Evans, T. van der Kwast, R. Allan, P. A. Humphrey, H. Grönberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Häkkinen, L. Egevad, M. Demkin, S. Dane, F. Tan, M. Valkonen, G. S. Corrado, L. Peng, C. H. Mermel, P. Ruusuvuori, G. Litjens, M. Eklund, A. Brilhante, A. Çakir, X. Farré, K. Geronatsiou, V. Molinié, G. Pereira, P. Roy, G. Saile, P. G. O. Salles, E. Schaafsma, J. Tschui, J. Billoch-Lima, E. M. Pereira, M. Zhou, S. He, S. Song, Q. Sun, H. Yoshihara, T. Yamaguchi, K. Ono, T. Shen, J. Ji, A. Roussel, K. Zhou, T. Chai, N. Weng, D. Grechka, M. V. Shugaev, R. Kiminya, V. Kovalev, D. Voynov, V. Malyshev, E. Lapo, M. Campos, N. Ota, S. Yamaoka, Y. Fujimoto, K. Yoshioka, J. Juvonen, M. Tukiainen, A. Karlsson, R. Guo, C.-L. Hsieh, I. Zubarev, H. S. T. Bukhar, W. Li, J. Li, W. Speier, C. Arnold, K. Kim, B. Bae, Y. W. Kim, H.-S. Lee, J. Park, and the PANDA challenge consortium (2022): "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge," *Nature Medicine*, 28, 154–163.

Campbell, A., P. E. Converse, W. E. Miller, and D. E. Stokes (1960): *The American Voter*, New York and London: John Wiley & Sons, Inc., 1 ed.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020): "Are Referees and Editors in Economics Gender Neutral?" *The Quarterly Journal of Economics*, 135, 269–327.

——— (2022): "Gender Differences in Peer Recognition by Economists," *Econometrica*, 90, 1937–1971.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024): "On Binscatter," *American Economic Review*, 114, 1488–1514.

Chen, Y. (2021): "Team-Specific Human Capital and Team Performance: Evidence from Doctors," *American Economic Review*, 111, 3923–62.

Chetty, R., D. J. Deming, and J. N. Friedman (2023): "Diversifying Society's Leaders? The

Causal Effects of Admission to Highly Selective Private Colleges," *National Bureau of Economic Research Working Paper*.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2020): "Income Segregation and Intergenerational Mobility Across Colleges in the United States," *The Quarterly Journal of Economics*, 135, 1567–1633.

Collins, W. J. and M. H. Wanamaker (2022): "African American Intergenerational Economic Mobility since 1880," *American Economic Journal: Applied Economics*, 14, 84–117.

Constantine, N. and S. Correia (2021): "Putting the "i" in Fixed Effects: Individual Fixed Effects with Group Level Outcomes," *mimeo Board of Governors of the Federal Reserve System*.

Cornelissen, T., C. Dustmann, and U. Schönberg (2017): "Peer Effects in the Workplace," *The American Economic Review*, 107, 425–456.

Croix, D. d. l. and M. Goñi (2024): "Nepotism vs. Intergenerational Transmission of Human Capital in Academia (1088-1800)," *Journal of Economic Growth*, 29, 469–514.

Crémer, J., L. Garicano, and A. Prat (2007): "Language and the Theory of the Firm," *The Quarterly Journal of Economics*, 122, 373–407.

Dal Bó, E., F. Finan, O. Folke, T. Persson, and J. Rickne (2017): "Who Becomes a Politician?" *The Quarterly Journal of Economics*, 132, 1877–1914.

Deming, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 132, 1593–1640.

Dessein, W. and T. Santos (2006): "Adaptive Organizations," *Journal of Political Economy*, 114, 956–995.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, ed. by J. Burstein, C. Doran, and T. Solorio, Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Dossi, G. (2024): "Race and Science," *mimeo UCL*.

Draca, M. and C. Schwarz (2024): "How Polarised are Citizens? Measuring Ideology from the Ground up," *The Economic Journal*, 134, 1950–1984.

Duch, R. M., H. d. Palmer, and C. J. Anderson (2000): "Heterogeneity in Perceptions of National Economic Conditions," *American Journal of Political Science*, 44, 635–652.

Duch, R. M. and R. T. Stevenson (2011): "Context and Economic Expectations: When Do Voters Get It Right?" *British Journal of Political Science*, 41, 1–31.

Eil, D. and J. M. Rao (2011): "The Good News–Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 3, 114–138.

Einio, E., J. Feng, and X. Jaravel (2022): "Social Push and the Direction of Innovation," *mimeo CEP LSE*.

ETS (2009): "GRE Guide to the Use of Scores 2009-2010, Extended Table 4," https://web.archive.org/web/20121222214014/http://ets.org/Media/Tests/GRE/pdf/gre_0910_guide_extended_table4.pdf, archived 2012/12/22, last retrieved 2025/03/07.

Fiorina, M. P. (1981): *Retrospective voting in American national elections*, New Haven: Yale Univ.Pr.

Freund, L. (2024): "Superstar Teams," *mimeo Columbia Business School*.

Frömmgen, A., J. Austin, P. Choy, N. Ghelani, L. Kharatyan, G. Surita, E. Khrapko, P. Lamblin, P.-A. Manzagol, M. Revaj, M. Tabachnyk, D. Tarlow, K. Villela,

D. Zheng, S. Chandra, and P. Maniatis (2024): "Resolving Code Review Comments with Machine Learning," in *2024 IEEE/ACM 46th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*.

Garicano, L. (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108, 874–904.

Garicano, L. and E. Rossi-Hansberg (2006): "Organization and Inequality in a Knowledge Economy," *The Quarterly Journal of Economics*, 121, 1383–1435.

Gentzkow, M. and J. M. Shapiro (2010): "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78, 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019): "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech," *Econometrica*, 87, 1307–1340.

Gino, F., M. I. Norton, and R. A. Weber (2016): "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives*, 30, 189–212.

Goldin, C. and L. F. Katz (2009): *The race between education and technology*, Harvard University Press.

Golman, R., D. Hagmann, and G. Loewenstein (2017): "Information Avoidance," *Journal of Economic Literature*, 55, 96–135.

Green, D. P. and B. Palmquist (1994): "How Stable Is Party Identification?" *Political Behavior*, 16, 437–466.

h2o.ai (2025): "Meet our Kaggle Grandmasters," https://h2o.ai/company/team/kaggle-grandmasters/, last retrieved 2025/03/07.

Hager, S., C. Schwarz, and F. Waldinger (2024): "Measuring Science: Performance Metrics and the Allocation of Talent," *American Economic Review*, 114, 4052–90.

Hamilton, B. H., J. A. Nickerson, and H. Owan (2003): "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111, 465–497.

He, S., R. Huang, J. Townley, R. C. Kretsch, T. G. Karagianes, D. B. T. Cox, H. Blair, D. Penzar, V. Vyaltsev, E. Aristova, A. Zinkevich, A. Bakulin, H. Sohn, D. Krstevski, T. Fukui, F. Tatematsu, Y. Uchida, D. Jang, J. S. Lee, R. Shieh, T. Ma, E. Martynov, M. V. Shugaev, H. S. T. Bukhari, K. Fujikawa, K. Onodera, C. Henkel, S. Ron, J. Romano, J. J. Nicol, G. P. Nye, Y. Wu, C. Choe, W. Reade, E. Participants, and R. Das (2024): "Ribonanza: deep learning of RNA structure through dual crowdsourcing," *bioRxiv : the preprint server for biology*.

Healy, A. and N. Malhotra (2013): "Retrospective Voting Reconsidered," *Annual Review of Political Science*, 16, 285–306.

Hengel, E. (2022): "Publishing while Female: Are Women Held to Higher Standards? Evidence from Peer Review," *The Economic Journal*, 132, 2951–2991.

Herkenhoff, K., J. Lise, G. Menzio, and G. M. Phillips (2024): "Production and Learning in Teams," *Econometrica*, 92, 467–504.

Hetherington, M. J. (1996): "The Media's Role in Forming Voters' National Economic Evaluations in 1992," *American Journal of Political Science*, 40, 372–395.

Hibbs, D. A., R. D. Rivers, and N. Vasilatos (1982): "The Dynamics of Political Support for American Presidents Among Occupational and Partisan Groups," *American Journal of Political Science*, 26, 312–332.

Hill, R. and C. Stein (2025): "Race to the Bottom: Competition and Quality in Science," *The*

Quarterly Journal of Economics, 140, 1111–1185.

Howard, J. (2024): "Is it a bird? Creating a model from your own data," https://www.kaggle.com/code/jhoward/is-it-a-bird-creating-a-model-from-your-own-data, last retrieved 2025/03/07.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019): "The Allocation of Talent and U.S. Economic Growth," *Econometrica*, 87, 1439–1474.

Iaria, A., C. Schwarz, and F. Waldinger (2018): "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science," *The Quarterly Journal of Economics*, 133, 927–991.

——— (2024): "Gender Gaps in Academia: Global Evidence Over the Twentieth Century," *mimeo LMU Munich*.

Ingram, P. (2021): "The forgotten dimension of diversity," *Harvard Business Review*, 99, 58–67.

IPUMS (2024a): "Census Occupation Codes, 1950 Basis," https://usa.ipums.org/usa-action/variables/OCC1950#codes_section, last retrieved 2024/09/07.

——— (2024b): "Integrated Occupation and Industry Codes and Occupational Standing Variables in the IPUMS," https://usa.ipums.org/usa/chapter4/chapter4.shtml, last retrieved 2024/09/07.

Jaravel, X., N. Petkova, and A. Bell (2018): "Team-Specific Capital and Innovation," *American Economic Review*, 108, 1034–73.

Jarosch, G., E. Oberfield, and E. Rossi-Hansberg (2021): "Learning from Coworkers," *Econometrica*, 89, 647–676.

Jones, B. F. (2009): "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *The Review of Economic Studies*, 76, 283–317.

Kaggle.com (2024a): "How to Use Kaggle," https://www.kaggle.com/docs/competitions, last retrieved 2025/03/07.

——— (2024b): "Kaggle Community Guidelines," https://www.kaggle.com/community-guidelines, last retrieved 2025/03/07.

——— (2024c): "Kaggle Rankings," https://www.kaggle.com/rankings?group=competitions, last retrieved 2025/03/07.

——— (2025): "Kaggle Progression System," https://www.kaggle.com/progression, last retrieved 2025/03/07.

Kireyev, P. (2020): "Markets for ideas: prize structure, entry limits, and the design of ideation contests," *The RAND Journal of Economics*, 51, 563–588.

Koch, B. J. and D. Peterson (2024): "From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution," *mimeo University of Chicago*.

Koffi, M. (2024): "Innovative Ideas and Gender Inequality," *mimeo University of Toronto*.

Koning, R., S. Samila, and J.-P. Ferguson (2021): "Who do We Invent for? Patents by Women Focus More on Women's Health, but Few Women Get to Invent," *Science*, 372, 1345–1348.

Kozlowski, D., V. Larivière, C. R. Sugimoto, and T. Monroe-White (2022): "Intersectional Inequalities in Science," *Proceedings of the National Academy of Sciences of the United States of America*, 119.

Larcinese, V., R. Puglisi, and J. M. Snyder (2011): "Partisan bias in economic news: Evidence on the agenda–setting behavior of U.S. newspapers," *Journal of Public Economics*, 95, 1178–1189.

Lazear, E. P. and S. Rosen (1981): "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89, 841–864.

Le Yaouanq, Y. (2023): "A model of voting with motivated beliefs," *Journal of Economic Behavior & Organization*, 213, 394–408.

Lemus, J. and G. Marshall (2021): "Dynamic Tournament Design: Evidence from Prediction Contests," *Journal of Political Economy*, 129, 383–420.

——— (2024): "Teamwork in Contests," *The Review of Economics and Statistics*, 1–45.

Lotka, A. J. (1926): "The Frequency Distribution of Scientific Productivity," *Journal of the Washington Academy of Sciences*, 16, 317–323.

Lowry, D. T. (2008): "Network Tv News Framing of Good Vs. Bad Economic News under Democrat and Republican Presidents: A Lexical Analysis of Political Bias," *Journalism and Mass Communication Quarterly*, 85, 483–498.

Mas, A. and E. Moretti (2009): "Peers at Work," *American Economic Review*, 99, 112–45.

Merton, R. K. (1957): "Priorities in Scientific Discovery: a Chapter in the Sociology of Science," *American Sociological Review*, 22, 635–659.

Michelman, V., J. Price, and S. D. Zimmerman (2022): "Old Boys' Clubs and Upward Mobility among the Educational Elite," *The Quarterly Journal of Economics*, 137, 845–909.

Minni, V. (2024): "Making the Invisible Hand Visible: Managers and the Allocation of Workers to Jobs," *mimeo Chicago Booth*.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018): *Foundations of Machine Learning*, MIT Press, second edition ed.

Moreira, D. and S. Pérez (2022): "Who Benefits from Meritocracy?" *National Bureau of Economic Research Working Paper*.

Morgan, A. C., N. LaBerge, D. B. Larremore, M. Galesic, J. E. Brand, and A. Clauset (2022): "Socioeconomic Roots of Academic Faculty," *Nature Human Behaviour*, 6, 1625–1633.

Moser, P. and S. Kim (2022): "Women in Science. Lessons from the Baby Boom," *mimeo NYU and University of Pennsylvania*.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Nobelprize.org (2024): "Nomination Archive," www.nobelprize.org/nomination/archive, last retrieved 2024/02/27.

Novosad, P., S. Asher, C. Farquharson, and E. Iljazi (2024): "Access to Opportunity in the Sciences: Evidence from the Nobel Laureates," *Centre for Economic Policy Research Discussion Paper*.

NVIDIA (2025): "Kaggle Grandmasters," https://www.nvidia.com/en-us/ai-data-science/kaggle-grandmasters/, last retrieved 2025/02/25.

Oster, E., I. Shoulson, and E. R. Dorsey (2013): "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease," *American Economic Review*, 103, 804–830.

Pavão, A. (2023): "Methodology for Design and Analysis of Machine Learning Competitions," Theses, Université Paris-Saclay.

Pearce, J. (2023): "Idea Production and Team Structure," *mimeo Federal Reserve Bank of New York*.

Pew Research Center (2016): "2016 Party Identification Detailed Tables," https://www.people-press.org/2016/09/13/2016-party-identification-detailed-tables/, last retrieved 2019/07/15.

——— (2018): "Party Identification Trends, 1992–2017," https://www.people-press.org/2018/03/20/party-identification-trends-1992-2017/, last retrieved 2019/07/15.

Plotts, J. and M. Risdal (2023): "Meta Kaggle Code," https://www.kaggle.com/ds/3240808, downloaded 2024/01/08.

Prior, M., G. Sood, and K. Khanna (2015): "You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions," *Quarterly Journal of Political Science*, 10, 489–518.

Reade, W. (2024): "Submission selections can only be made by team leads," https://www.kaggle.com/competitions/home-credit-default-risk/discussion/64440, last retrieved 2025/03/07.

Reich, V. (2023): "Signal or Noise - Signalling Skill among Data Professionals," *mimeo University of Munich*.

Risdal, M. and T. Bozsolik (2022): "Meta Kaggle," https://www.kaggle.com/ds/9, downloaded 2024/2/07.

Roelofs, R., V. Shankar, B. Recht, S. Fridovich-Keil, M. Hardt, J. Miller, and L. Schmidt (2019): "A Meta-Analysis of Overfitting in Machine Learning," in *Advances in Neural Information Processing Systems*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Curran Associates, Inc., vol. 32.

Ross, M., B. Glennon, R. Murciano-Goroff, E. Berkes, B. Weinberg, and J. Lane (2022): "Women are credited less in science than men," *Nature*, 608, 135–145.

Rossiter, M. W. (1982): *Women Scientists in America: Struggles and Strategies to 1940*, vol. 1, JHU Press.

——— (1998): *Women scientists in America: Before affirmative action, 1940-1972*, vol. 2, JHU Press.

Ruggles, S., C. Fitch, R. Goeken, J. D. Hacker, J. Helgertz, E. Roberts, M. Sobek, K. Thompson, J. R. Warren, and J. Wellington (2019): "IPUMS Multigenerational Longitudinal Panel," .

Ruggles, S., C. A. Fitch, R. Goeken, J. D. Hacker, M. A. Nelson, E. Roberts, M. Schouwiler, and M. Sobek (2024): "IPUMS Ancestry Full Count Dataset: Version 4.0," .

Schaffner, B. F. and C. Roche (2017): "Misinformation and Motivated Reasoning: Responses to Economic News in a Politicized Environment," *Public Opinion Quarterly*, 81, 86–110.

Smith, A. (1776): "An Inquiry into the Nature and Causes of the Wealth of Nations," available through Project Gutenberg, https://www.gutenberg.org/ebooks/38194, last retrieved 2025/03/07.

Stansbury, A. and K. Rodriguez (2024): "The Class Gap in Career Progression: Evidence from US Academia," *mimeo Massachusetts Institute of Technology*.

Stansbury, A. and R. Schultz (2023): "The Economics Profession's Socioeconomic Diversity Problem," *Journal of Economic Perspectives*, 37, 207–230.

Stantcheva, S. (2020): "Understanding Economic Policies: What do people know and learn?" *mimeo Harvard University*.

Sun, L. and S. Abraham (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225, 175–199.

Taylor, C. R. (1995): "Digging for Golden Carrots: An Analysis of Research Tournaments," *The American Economic Review*, 85, 872–890.

Teodoridis, F. (2018): "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise," *Management Science*, 64, 3625–3648.

Teodoridis, F., M. Bikard, and K. Vakili (2019): "Creativity at the Knowledge Frontier: The Impact of Specialization in Fast- and Slow-paced Domains," *Administrative Science Quarterly*, 64, 894–927.

Thorp, H. H. (2023): "It Matters Who Does Science," *Science*, 380, 873–873.

Trotman, J. (2019): "Meta Kaggle: Competition Shake-up," https://www.kaggle.com/jtrotman/meta-kaggle-competition-shake-up, last retrieved 2025/03/07.

Truffa, F. and A. Wong (2022): "Undergraduate Gender Diversity and Direction of Scientific Research," *mimeo Northwestern University*.

U.S. Census Bureau (2018): "2013–2017 American Community Survey 5–Year Estimates," http://factfinder.census.gov, last retrieved 2019/07/13.

Uzzi, B., S. Mukherjee, M. Stringer, and B. Jones (2013): "Atypical Combinations and Scientific Impact," *Science*, 342, 468–472.

van Leeuwen, M. and I. Maas (2011): *Hisclass: A Historical International Social Class Scheme*, G - Reference, Information and Interdisciplinary Subjects Series, Leuven University Press.

Waldinger, F. (2011): "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *The Review of Economic Studies*, 79, 838–861.

Wang, Y., L. Hung, A. D. Gotmare, N. D. Q. Bui, J. Li, and S. C. H. Hoi (2023): "CodeT5+: Open Code Large Language Models for Code Understanding and Generation," *mimeo Salesforce AI Research*.

Weidmann, B. and D. J. Deming (2021): "Team Players: How Social Skills Improve Team Performance," *Econometrica*, 89, 2637–2657.

Weidmann, B., J. Vecci, F. Said, D. J. Deming, and S. R. Bhalotra (2024): "How Do You Find a Good Manager?" *National Bureau of Economic Research Working Paper*.

Westwood, S. J., S. Iyengar, S. Walgrave, R. Leonisio, L. Miller, and O. Strijbis (2018): "The tie that divides: Cross–national evidence of the primacy of partyism," *European Journal of Political Research*, 57, 333–354.

Williams, L. A. and M. Y. Bartlett (2015): "Warm thanks: gratitude expression facilitates social affiliation in new relationships via perceived warmth," *Emotion*, 15, 1–5.

Wolfers, J. (2007): "Are Voters Rational? Evidence from Gubernational Elections," *mimeo University of Michigan*.

Wright, F. A. and A. A. Wright (2018): "How surprising was Trump's victory? Evaluations of the 2016 U.S. presidential election and a new poll aggregation model," *Electoral Studies*, 54, 81–89.

Wuchty, S., B. F. Jones, and B. Uzzi (2007): "The Increasing Dominance of Teams in Production of Knowledge," *Science*, 316, 1036–1039.

Zimmermann, F. (2020): "The Dynamics of Motivated Beliefs," *American Economic Review*, 110, 337–361.

Statement on Used Generative Models

ChatGPT 3.5 (OpenAI, https://chat.openai.com)
Gemini (Google, https://gemini.google.com/)
were used in the preparation of this dissertation for proofreading and data cleaning support.