I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18348

Who Studies What? Country of Origin, Gender, and Field Specialization among Economics PhDs

Karan Singhal
Eva Sierminska

DECEMBER 2025

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18348

Who Studies What? Country of Origin, Gender, and Field Specialization among Economics PhDs

Karan Singhal
University of Luxembourg and LISER

Eva Sierminska
LISER, INE PAN and IZA

DECEMBER 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Who Studies What? Country of Origin, Gender, and Field Specialization among Economics PhDs

We study the determinants of field specialization among U.S. economics PhD students, focusing on individual, institutional, and contextual factors shaping early research careers. Using data on over 8,000 dissertations from 2009–2018, we classify each dissertation into one of ten fields using author-reported JEL codes and topic modeling of abstracts. We link dissertations to student gender, program characteristics, and country of origin inferred from surnames and matched to country-level indicators. We find substantial variation in field choice by region of origin. Gender gaps in specialization are not uniform but vary in size and direction across regions, indicating that gender and origin interact in shaping choices. Results are robust to alternative classification methods and to using genetic distance as a continuous measure of origin. Our findings highlight how early specialization in economics reflects inherited context and institutional exposure, with implications for research agendas, job market outcomes, and diversity across subfields.

JEL Classification: J15, J16

Keywords: field specialization, economics PhD students, JEL, topic modeling, diversity in economics

Corresponding author:
Eva Sierminska
LISER
11 Porte de Sciences
4366 Esch/Alzette
Luxembourg

E-mail: eva.sierminska@liser.lu

1 Introduction

Persistent demographic imbalances continue to characterize the modern economics profession. The literature on inequality within economics academia, summarized in Singhal and Sierminska, 2025, documents numerous disparities across the educational and professional pipeline, as well as, biases and discrimination faced by underrepresented groups. Highlighted is the fact that women's share of doctorates and faculty positions has hovered around one-third since the early 2000s, with little evidence of sustained progress (Lundberg and Stearns, 2019; Bayer et al., 2020). Racial and ethnic underrepresentation is even more stark: Black, Hispanic, and Native American scholars constitute a small fraction of U.S. PhD recipients and an even smaller share of tenured faculty and journal editors. These demographic gaps are further compounded by socio-economic stratification. Economics disproportionately includes students whose parents hold advanced degrees or occupy high-income professions, making it the least socio-economically diverse academic discipline in the United States (Stansbury and Schulz 2023). Taken together, these figures remain seriously below the population shares of women and minority groups in the United States, underscoring the extent of underrepresentation in the profession.

These compositional imbalances constrain the discipline's intellectual agenda. Advani et al. (2020, 2021) show that fewer than two percent of economics publications since 1970 have addressed race-related topics (compared to four percent in political science and twelve percent in sociology) and that this limited body of work is disproportionately authored by minority economists. Homogeneity may also dampen scientific creativity: ethnically diverse research teams tend to produce more novel and highly cited work (Freeman and Huang, 2014). Male and female economists differ systematically in policy views on minimum wages, labor standards, healthcare, and inequality (May et al., 2014), and women are more likely to pursue interdisciplinary and applied research topics (Bayer and Rouse, 2016; Bayer et al., 2020). Underrepresentation thus risks narrowing both the range of questions posed and the types of evidence that influence public policy.[1] Progress will remain slow unless the discipline understands better where in the pipeline disparities emerge and how they channel different groups into (or away from) particular research areas. This paper focuses on one such channel: field specialization choice (such as microeconomics, macroeconomics, labor economics) in graduate school, an understudied but consequential step in shaping economists' trajectories.

We document and examine how disparities in field specialization manifest at the doctoral stage by drawing on data from multiple sources to analyze over 8000 dissertations completed in U.S. economics PhD programs between 2009 and 2018. Each dissertation is assigned to one of ten broad fields using two sources. First, the candidates' self-reported Journal of Economic Literature (JEL) codes are mapped into fields and then second, cross-validated using unsupervised topic modeling estimated on more than 29,000 abstracts. This approach helps address multi-field specializations and mitigates potential misclassifications (Ambrosino et al., 2018; Blei et al., 2003). Author gender is inferred from given names, while country of origin is inferred using surname-based matches to global genealogy data, and then aggregated into seven world

[1] Yellen (2019) and Mester (2019) further argue that a more diverse profession is better equipped to diagnose economic problems affecting heterogeneous populations, while Witteman et al. (2021) emphasize how scholars' lived experiences can sharpen policy insight.

regions.[2] To further potentially capture the institutional and cultural context of the country of origin, we supplement the regional classification with a small set of country-level markers (colonial heritage, Muslim population share, and former-USSR status). These variables provide a way to probe systematic legacies that region categories alone may mask, while also serving as parsimonious proxies where country-of-origin fixed effects are infeasible.

Using multinomial logistic and linear regression models, three main findings emerge. First, region of origin is a strong predictor of field choice - for instance, students from Sub-Saharan Africa are significantly more likely to specialize in Development, while East Asian origin students have higher representation in Econometrics. Second, gender gaps in field specialization are not uniform across regions. For example, gender gaps in fields like Econometrics and Macro/Finance vary sharply by region, with male overrepresentation in these fields more pronounced in only some regions. These differences reveal that gendered sorting into fields reflects not only gender, but its interaction with region-specific contexts. Third, the drivers of field specialization differ across regions, indicating that students' choices are shaped by a combination of context-specific influences (such as institutional rank or country background) rather than a single, uniform pattern.

Field specialization often takes shape early in the career of an economist, typically during the PhD. The initial choice carries long-term consequences. Fortin et al. (2021) show that field composition explains a substantial share of the gender gap in academic placements outside the top-fifty departments. Certain fields may have denser citation networks, and greater funding opportunities; others offer less visibility and lower journal prestige. Systematic sorting by demographic characteristics can therefore amplify early disparities in placement, publication, and tenure outcomes. Using dissertation data, Sierminska and Oaxaca (2021, 2022) document that female doctoral students are significantly less likely than their male peers to specialize in macroeconomics/finance, industrial organization, public economics, or development, and more likely to choose agricultural or environmental economics. Conference programs tell a similar story: women are underrepresented in macro-finance sessions and overrepresented in applied microeconomics (Chari and Pinksmith-Goldham 2022; Beneito et al. 2021). Publication patterns also reflect these gaps. Male authors dominate theory and Macroeconomics journals, while female authors appear more frequently in development outlets (Onder and Yilnazkuday 2020). By contrast, evidence on the relationship between country or ethnic origin and field choice remains limited. Recent work by Antman et al. (2024) shows that underrepresented minorities (identified based on ethnicity and race) specialize in a handful of fields but are not especially likely to study race itself after controlling for characteristics. Outside of economics, Kozlowski et al. (2022a) analyze millions of scientific papers and find that the intersection of race and gender is an even stronger predictor of research topics than either dimension alone, highlighting the importance of multi-dimensional analyses.

Our work contributes to the literature in three ways. First, by examining how country and region of origin shape field selection. Our data reveal systematic clustering of non-U.S. origin (or foreign born) PhD economists across subfields - patterns that earlier work, which

[2] Although our method cannot distinguish first- from second-generation (or older) migrants, it remains informative regarding the inherited language, curricular background, and pre-doctoral education.

concentrates mainly on gender and, more recently, on underrepresented minority scholars, has largely overlooked. Country of origin may influence access to academic experiences, exposure to disciplinary norms, and the mentoring networks that steer career trajectories. It also likely reflects the influence of dominant ethnic groups and prevailing social norms, as well as historical economic factors transmitted across generations. These forces may shape both field and methodological choices in distinct ways, and may do so differently for women and men, contributing to variation in gender gaps across contexts (Jayachandran 2021; Zafar 2013). Second, by interacting the country of origin with gender, the analysis reveals that gender disparities in field specialization vary markedly across regions and subfields. These findings reveal the need for a more disaggregated approach to understanding and addressing gender disparities in the discipline. Third, it improves field classification by integrating self-reported JEL codes with text analysis of abstracts (specifically using Latent Dirichlet Allocation (LDA)), offering an alternative validation method and helping in identifying dominant fields. Prior analyses have largely relied on JEL codes alone; as Ambrosino et al. (2018) note, such tags may obscure the true content of research when used strategically for signaling.

The remainder of the paper is organized as follows. Section 2 describes the data sources, methods for assigning demographic characteristics and fields, and validation checks. Section 3 outlines the sample construction, summary statistics, and patterns in field specialization by gender and origin. Section 4 presents the main regression results and robustness checks, including alternative field definitions and country-of-origin measures. Section 5 concludes with implications and directions for future research.

2 Data Sources and Concepts

We compile data on over 8,000 individuals who earned economics doctorates from U.S. universities between 2009 and 2018, drawing on three complementary sources. First, *EconLit* provides each graduate's name, dissertation title, year of completion, awarding institution, and JEL codes. Second, *ProQuest* supplies dissertation titles and abstracts, and in most cases allows us to confirm the year of graduation and awarding institution. We merge these two datasets using a combination of fuzzy and manual matching on name, university, and cohort, successfully matching about 75% of students[3] listed in *EconLit*. Third, following Sierminska and Oaxaca (2021), we use the contemporaneous share of female faculty in each economics department. We supplement this with additional university-level information, including whether the institution is public or private and its ranking tier based on *U.S. News & World Report*.

The resulting dataset assigns to each graduate a consistent set of personal attributes (gender, region of origin, graduation year) and graduate-school environment variables (institutional rank, sector, and female-faculty shares).

[3] We compare the characteristics of the matched and unmatched samples and find them statistically similar in terms of gender composition and regional representation. However, the merged sample includes slightly older cohorts on average (by 0.4 years) and draws more heavily from higher-ranked institutions (indicating lower probabilty of those from lower ranked institutions uploading their theses on Proquest) - though representation from top 20 programs is comparable across both groups.

2.1 Gender Identification

We first apply a name-based algorithm using *Genderize.io*; names with a female probability of at least 0.80 are classified as female. Ambiguous cases (e.g., “Ali,” “Andrea,” “Soumya”) are resolved manually using CVs, pronouns, or photographs, following a similar procedure to Sierminska and Oaxaca (2022). While this approach cannot identify non-binary individuals, it aligns with the binary gender classifications commonly embedded in academic hiring, promotion, and funding processes.

2.2 Country of Origin

Country of origin is inferred primarily from surnames, with first names used for disambiguation in certain cases. We rely on *Forebears.io*, a global genealogy database that estimates the likelihood of a name originating from a particular country. This is based on two core metrics: the *density* of a name within a country (its relative popularity) and its *frequency* (the total number of recorded occurrences). These probabilities allow us to assign a national origin to each graduate in a systematic and scalable way.

Surname-based classification reflects ancestral heritage and perceived ethnicity, rather than nationality or citizenship. Names can signal ethnic identity across multiple generations and have been shown to influence both self-perception and external treatment (Biavaschi et al. 2017; Kerr and Kerr 2018). A third-generation Korean American, for example, may still be shaped by familial narratives that value quantitative aptitude, nudging her toward a field like econometrics. Although our method cannot distinguish between first- and second-generation migrants, it provides insight into patterns of possible cultural transmission that are difficult to observe through administrative records alone.

Validation

To validate the approach, we conducted a manual audit of 1,000 CVs. We collected detailed information on each economist’s stated nationality, citizenship, native language, and undergraduate institution, based on whatever was available. Based on patterns observed in this exercise, we developed a set of validation rules to improve the accuracy of our algorithm. For example, surnames are assigned to the country where they are most prevalent, provided the probability exceeds 60 percent. (This allowed for a complete match.)

However, there are some exceptions In cases where the match is the United States, we impose a stricter 80 percent threshold, and we override a U.S. assignment if the second-ranked country exceeds 30 percent. This adjustment is especially important for surnames that are common across the Anglo world, such as “Smith” or “Murphy.” Similarly, for Hispanic surnames that appear across multiple countries, we apply higher thresholds to separate closely related groups - for instance, distinguishing Brazilian from Portuguese origin.[4] We provide additional details on classification rules, probability thresholds, and tie-breaking logic in Note A1 in the Appendix.

[4] In ambiguous cases, we supplement these steps with a large language model prompt using Chat GPT with search-enabled capabilities. This allows us to go beyond name-based inference and retrieve biographical information about individual economists where available online. Where the large language model (LLM)-generated assignment aligns with the Forebears result, we automatically accept the classification. This combination of algorithmic inference, manual auditing, and AI-enabled search substantially increases our confidence in the assignments, particularly for non-Western names. See Note A1 in the Appendix for more details.

Despite its importance, we are unable to determine or accurately predict race. Surname-based methods are less reliable for American economists whose names are common across racial groups; for example, “Washington” or “Andrews” are associated with both White and Black populations. While surname origin correlates more strongly with race in other regions, for example, those bearing surnames with origins predicted as Sub-Saharan Africa, this is not uniformly true. For this reason, our analysis focuses on the region of origin rather than on race.

Name-based methods can introduce bias but, when carefully calibrated, they offer a practical way to uncover broad regional and ethnic patterns (Kozlowski et. al., 2022b). Our focus on the region of origin mitigates concerns around fine-grained misclassification and avoids overinterpreting ambiguous cases. Further, Antman et al. (2025) show that self-identification can shift across contexts and generations, particularly among later-generation immigrants. For example, many U.S.-born descendants of Mexican immigrants might not identify as Hispanic, often reflecting shifting social incentives or assimilation pressures. In contrast, surnames tend to provide a more stable signal of inherited cultural background and frequently shape how individuals are perceived externally, even when self-identification evolves. Of course, name changes - whether through marriage, assimilation, or personal choice - can weaken this link, but in practice surnames still function as a useful proxy for inherited identity and external perceptions. Similarly, name-based methods may still fall short in capturing multiracial or multiethnic identities and can be less reliable for identifying people of color. Yet recent work has emphasized the value of studying these patterns despite imperfect classification (Hofstra et al., 2022).

Aggregation to regions

Each inferred country of origin is assigned to one of seven world regions. These include North America (United States and Canada)[5]; Europe (both east and west); East Asia; South and Central Asia (South Asia); Latin America and the Caribbean (LAC/ South America); the Middle East, North Africa and West Asia (MENA); and Sub-Saharan Africa. This classification balances geographic proximity with (relative) institutional similarity and ensures that our categories are broad enough to avoid false precision while still capturing meaningful variation. A full list of countries and their assigned regions is provided in Table A1 in the Appendix.

The largest shares of PhD graduates in the sample originate from East Asia, followed by North America, South and Central Asia, and LAC. Gender patterns, however, reveal notable differences in regional composition. The main differences are that women are more likely than men to come from South and Central Asia (13% vs. 8%). In contrast, men are overrepresented among those from the North America, making up 30 percent of male graduates versus 23 percent of females.

[5] This also includes Australia and New Zealand for parsimony, given the limited number of surnames uniquely predicted to originate from these countries.

Table 1: PhD Graduates by Region of Origin

Region of Origin	**Males (%)**	**Females (%)**	**Overall (%)**
North America	29.88	22.81	27.81
Europe	15.00	16.92	16.00
LAC	8.31	6.04	7.65
South Asia / Central Asia	7.63	13.13	9.23
East Asia	31.76	32.88	32.00
Sub-Saharan Africa	2.84	2.56	2.76
MENA	4.68	5.65	4.96
Total	100	100	100
N	6,266	2,582	8,848

Note: Entries are column percentages by gender and for the full sample. The bottom row reports counts.

2.2.1 Genetic Distance

To complement the regional categorization, we incorporate a continuous measure of genetic distance to the United States, based on Spolaore and Wacziarg (2009).[6] Using this measure allows us to exploit information at the country-of-origin level - rather than only relying on broad regional groupings - and to test whether field specialization varies along a continuous gradient rather than at discrete regional boundaries. The index captures the expected time, in millennia, since two populations last shared common ancestors. Although framed in biological terms, genetic distance closely tracks long-run cultural proximity: countries in Western Europe, whose populations share relatively recent ancestry with the U.S., are genetically closest, while Sub-Saharan Africa, East Asia, and parts of South and Central Asia lie furthest away.

While genetic distance inevitably overlaps with visible ancestry (and by extension, race), it does not map neatly onto contemporary racial categories. Its value lies in the structure it imposes - as prior research shows that genetic distance correlates with shared cultural, linguistic, and institutional traits (Spolaore and Wacziarg 2009; 2015), making it a useful proxy for historically transmitted influences. These may be experienced directly or inherited intergenerationally through family norms and educational systems. In this way, genetic distance complements the regional classification by helping to identify whether field specialization patterns reflect sharp breaks across regions or a more continuous gradient.

Table 2 presents average genetic distance to the United States by region of origin, calculated by assigning each individual the genetic distance score of their inferred country of origin and aggregating across regions. By construction, U.S. and Canadian graduates have a genetic distance of zero. As expected, Europe follows as the closest region, while East Asia and Sub-Saharan Africa exhibit the largest average distances.

We assign each inferred country of origin its corresponding genetic distance score and estimate alternative specifications that replace regional dummies with this continuous variable. The goal is to test whether the probability of specializing in specific fields varies systematically

[6] This approach seems particularly useful, as it focuses directly on the intersection of genetic distance and economic outcomes, making it particularly relevant for examining how inherited context might influence research trajectories among PhD students from varying regions. The range of the measure is 0 to 1.

Table 2: Average Genetic Distance of PhD Graduates' Region of Origin to USA

Region of Origin	**Mean**	**Min**	**Max**	**SD**	**Median**
North America	0.000	0.000	0.021	0.001	0.000
Europe	0.013	0.010	0.017	0.003	0.015
MENA	0.019	0.015	0.041	0.004	0.016
South Asia / Central Asia	0.021	0.016	0.039	0.003	0.021
Latin & S. America (LAC)	0.028	0.013	0.048	0.011	0.036
East Asia	0.041	0.037	0.044	0.002	0.040
Sub-Saharan Africa	0.046	0.027	0.052	0.005	0.048

Note: Mean calculated based on average distance from country of origin and then aggregated to regions. SD - standard deviation

along a broader gradient of inherited proximity to U.S. roots.

2.3 Field Classification

2.3.1 JEL Codes

Every dissertation indexed in EconLit reports up to seven JEL codes. Following Sierminska and Oaxaca (2022), we group these codes into ten broad field categories. If a dissertation's modal JEL codes are within a single category, we assign that category as the dissertation's stated or dominant field. Approximately 85 percent of dissertations meet this criterion (7,496 out of 8,848).

Table 3, shows the details of the JEL codes' classification into ten categories for brevity and analytical feasibilites. Health, Education and Welfare (I) and Labor and Demographic Economics (J) are clubbed; so are Macroeconomics (E) and Financial Economics (G); and Economic Development, Innovation, Technological Change (O) and International Economics (F). We follow these categories in the paper, and the main regressions.

2.3.2 Topic Modelling of Abstracts

To complement the structured JEL-based classification, we apply topic modeling to dissertation abstracts using Latent Dirichlet Allocation (LDA) - a machine learning method that scans large volumes of text and groups together words that tend to appear in similar contexts. In simple terms, LDA helps us identify the main themes or topics that each dissertation is about, based on the words used in the abstract. Our model is estimated on 29,000 abstracts retrieved from ProQuest for the years 2000 to 2021.[7] Before running the model, we clean the text by removing common words that are extremely common or carry limited meaning for identification (e.g., "the", "this", "model") and technical filler words specific to economics. We also apply stemming, which means reducing words to their base form (for example, treating "estimating" and "estimation" as the same word).

After testing different versions of the model with varying numbers of topics and settings, we settle on ten topics, which strike a balance between being statistically coherent (words within

[7] A subset of these abstracts is matched to the main sample (of 8000+ observations) drawn from EconLit, which includes JEL classification codes. Using a combination of name, university, and graduation year, we successfully match approximately 75% of the EconLit sample.

Table 3: Classification of JEL Codes

Field	JEL Codes	Detailed JEL Codes
Econometrics	C	C. Mathematical and Quantitative Methods
Micro	D	D. Microeconomics
Labor	I, J	I. Health, Education and Welfare; J. Labor and Demographic Economics
Macro/Finance	E, G	E. Macroeconomics and Monetary Economics; G. Financial Economics
IO	L	L. Industrial Organization
Environmental & Agricultural	Q	Q. Agricultural & Natural Resource Economics; Environmental Economics
Public	H	H. Public Economics
Development/Growth/International	F, O	O. Economic Development, Innovation, Technological Change, and Growth; F. International Economics
Economic History	B, N	B. History of Economic Thought, Methodology, and Heterodox Approaches; N. Economic History
Other	P, A, K, M, R, Y, Z	P. Economic Systems; A. General Economics and Teaching; Z. Other Topics; K. Law and Economics; R. Urban, Regional, Real Estate & Transportation Economics

a topic make sense together) and easily understandable. This number also aligns with the ten broad fields we use from the JEL codes, making it easier to compare the two classification methods (Ambrosino et al. 2018). See Table 4 for details.

The model assigns each abstract a probability distribution across the ten topics, with weights summing to one. We treat the topic with the highest weight as the abstract's dominant theme. To improve interpretability, we use a large language model (ChatGPT) to generate concise labels based on the top 20 keywords associated with each topic. Table 4 presents these topic labels alongside their defining keywords.

These are then aligned manually with the ten JEL groups.[8]

When compared to the 85 percent of dissertations with single JEL fields, the broader ordering of specializations is broadly consistent across both classifications (e.g.- 80%+ identified as labor economics completely align, and so on). While some divergences emerge, the overarching patterns - such as the empirical/theoretical divide and the prominence of Labour/Education and Macroeconomics and other ordering of fields - remain qualitatively similar.[9] Appendix Table

[8] The alignment between topic labels and field classifications is validated by comparing them to JEL codes across the sample. In six of the ten cases, topic groups overlap with the original JEL categories – Labor, Econometrics, Macro, Agriculture, Econometrics, Development. The remaining four blend elements from multiple fields and are grouped accordingly. For instance, Health, Education, and Labor appear as distinct topics but are combined into one field – "Labor" (following JEL classifications I and J); similarly, separate topics for Macroeconomics and Finance are grouped under "Macroeconomics and Finance" to match the broader classification scheme. As a robustness and sensitivity check, we use the original topic probabilities across all ten topics as continuous outcomes in Table 10. The main takeaways are largely consistent.

[9] There are some differences worth noting such as the incidence of broad categories like Microeconomics and Public Economics being less frequently identified by the topic model. The output reveals that many of these theses are better represented as focusing on Econometrics or Industrial Organization instead, with hints of Public

Table 4: Topic/Field Assignment based on Topic Modeling Keyword Output

Keywords	**Topic Assigned**
“health”, “insur”, “care”, “women”, “cost”, “children”, “patient”	Health
“econom”, “develop”, “social”, “institut”, “polit”, “commun”	Development
“price”, “food”, “agriculture”, “water”, “demand”, “consumer”, “farmer”	Agriculture/Environment
“rate”, “market”, “financi”, “price”, “bank”, “risk”, “monetari”	Finance
“firm”, “inform”, “market”, “decis”, “cost”, “competit”, “consum”, “behavior”, “contract”, “incentive”	IO / Game Theory
“trade”, “product”, “export”, “import”, “invest”, “polici”	Trade
“labor”, “household”, “wage”, “worker”, “income”, “employment”	Labor
“school”, “educ”, “student”, “hous”, “colleg”, “program”	Education
“tax”, “govern”, “polici”, “income”, “public”, “polit”, “econom”	Macroeconomics
“estim”, “model”, “distribut”, “gener”, “propos”, “equilibrium”, “optim”	Econometrics

Note: For comparability with disciplinary classifications, these ten topics are mapped to six of the ten major JEL field categories: (i) Labor (combining Labor, Health, and Education topics), (ii) Development and Growth (combining Development and Trade topics), (iii) Agriculture/Environment, (iv) Macro/Finance (combining Finance and Macroeconomics topics), (v) Econometrics, and (vi) IO (Industrial Organization / Game Theory). The “Macroeconomics” topic identified by the model also contains frequent references to tax, govern, and public, which conceptually overlap with Public Economics. Hence, this topic may be viewed as encompassing both macroeconomic and fiscal-policy–related work. This harmonization ensures consistency between the topic-model output and the broader JEL framework while retaining substantive distinctions across related areas. The original ten topic-based classifications are also retained for the alternate regression specification described in Section 3.

A2 provides a detailed comparison of JEL and topic-modeling-based classifications, illustrating their overall coherence and the few areas of divergence.

2.3.3 Combining JEL and Topic Modeling output

To harmonize field classification, we combine JEL codes with topic modelling. If a dissertation’s self-reported JEL codes fall clearly within a single broad category, we retain that category as the assigned field. For dissertations where JEL codes are missing or span multiple categories, we assign the field based on the dominant topic identified by the Latent Dirichlet Allocation (LDA) model. For instance, if a dissertation is tagged with both Labor and Macroeconomics JEL codes but the LDA assigns a higher probability to Labor, the final classification is Labor. In a small number of cases, topic labels conceptually overlapped (e.g., Industrial Organization, Econometrics, and Microeconomics). For consistency, we applied simple decision rules to assign these to the broader category they most closely align with (e.g., Industrial Organization as part of Microeconomics).[10]

After applying these classification rules, we arrive at a harmonized dataset of single-field dissertations used in the main analysis. In this way, we gained 689 observation giving us a total of 8,185 (7,496 + 689). Table 5 presents the final distribution across broad field categories. Labor/Health (I/J) emerges as the most common specialization, accounting for nearly

Economics also captured under topics labeled as Macroeconomics (through fiscal and policy terms) and elements of Micro appearing within the IO/Game Theory cluster.

[10] To validate the field assignments in ambiguous cases, we relied on a combination of manual and AI-assisted review. We used ChatGPT’s reasoning model to assist with interpretation of abstract content and topic-label fit, particularly in edge cases. In addition, two external evaluators (PhD economics students) were asked to review subsets of classifications to ensure conceptual consistency across overlapping fields.

25 percent of the sample, followed by Macroeconomics and Finance at 19 percent. The remaining dissertations are spread across other fields, including Microeconomics, Development, and Econometrics.

Table 5: Dominant Field Specialization (based on a combination of Topic Modeling and JEL codes)

Field	**Share (%)**
Econometrics (C)	5.88
Micro (D)	12.71
Labor/Health (I,J)	24.62
Macro/Finance (E,G)	19.35
IO (L)	5.95
Environ & Agric (Q)	9.26
Public (H)	2.41
Dev/Growth/Int (O,F)	12.96
Econ History (B,N)	0.99
Others (P, A, K, M, R, Y, Z)	5.88
Total	100.00
N	8,185

2.4 Country-level Factors

While regional classifications capture groups of neighboring countries that often share socio-economic and cultural features, they may mask meaningful within region variation in national institutions (and the norms embedded within them) that shape educational trajectories and field choices. To account for such within-region heterogeneity, we augment the dataset with a set of country-specific historical and religious variables. These are not intended to represent structural channels in a formal causal sense but rather serve as proxies for inherited institutional templates that precede migration and may persist across generations through socialization and norms.

Specifically, we incorporate three dimensions of national context: colonial legacy, affiliation with the former Soviet Union, and religious composition (specifically the share of Muslim population).[11]

These country-level characteristics are used in two ways. First, they are explored as potential determinants of field specialization, allowing us to test whether inherited national models such as colonial governance structures, exposure to Soviet academic traditions, or prevailing religious norms systematically influence students' academic trajectories. Second, they are included as controls to help isolate the effect of region of origin. Since we cannot include country fixed

[11] We avoid incorporating contemporary country-level indicators such as GDP, HDI, Gini coefficient, gender inequality indices, education indices (e.g., share of STEM graduates, PISA scores), or R&D expenditure, as these variables vary over time and would require more temporal alignment with each individual's formative years. Since our dataset does not distinguish whether international graduates were nationals, long-term residents, or second- or third-generation migrants, applying such indicators uniformly may therefore more likely to misrepresent the context shaping individual trajectories. Exploring the interaction between the historical factors and more recent, contemporaneous country characteristics (e.g., conditions at time of PhD enrollment) would be a valuable direction for future research but remains outside the present scope.

effects (due to the potential errors in classification and limited observations within each country of origin), these variables serve as proxy to account for some variation within each region. In doing so, they help reduce regional indicators from capturing unobserved institutional or cultural aspects specific to individual countries, and help to distinguish whether observed patterns stem from broad regional proximity or other other influences.

2.4.1 Colonial Legacy

Colonial rule left lasting marks on the legal, educational, and administrative foundations of many countries, with effects that extend well beyond formal independence. As Acemoglu et al. (2001) and Sokoloff and Engerman (2000) note, colonial governance shaped not only legal and property rights systems but also the design of curricula, the language of instruction, and the organization of civil services. These inherited structures may continue to shape educational and occupational pathways - not only for foreign students educated outside the U.S., but also for second- and third-generation students within the U.S., through the intergenerational transmission of norms and institutional influences experienced by their parents or grandparents (Ferrara and Luthra 2024).

To capture these institutional legacies, we classify countries based on whether they were formerly colonized by Britain, France, Spain, or Portugal.[12] This draws on existing distinctions between settler and extractive colonial regimes and the institutional frameworks they left behind (Acemoglu et al. 2001; Banerjee and Iyer 2005). Our focus is on countries governed under more extractive colonial arrangements, where bureaucratic, legal, and/or educational systems endured well beyond independence, shaping not only state institutions but also cultural norms and expectations around higher education.

We construct four mutually exclusive indicators for countries with dominant British, French, Spanish, or Portuguese colonial influence.[13] These four empires governed much of the non-European world between the sixteenth and twentieth centuries. While colonialism is not the only force shaping modern institutions, it serves as a proxy for inherited national models, particularly those that influenced the academic and professional environments of earlier generations. In cases of overlap, where multiple colonial powers were present (such as British and Portuguese rule in India), countries are classified based on the power that most durably shaped their post-independence institutions.[14]

A full list of countries included in each colonial category is provided in Table A3, with illus-

[12] We exclude countries with Dutch colonial heritage because their colonial structures - typically commercial, maritime, and indirect governance - differ from the more extractive legal-bureaucratic frameworks imposed by Britain, France, Spain, and Portugal, which are the focus of this chapter.

[13] *Settler* colonies such as the United States, Canada, Australia, and New Zealand, are not coded as British. These countries, while initially under British rule, evolved rapidly into self-governing states with institutional trajectories that diverged significantly from the extractive colonial template. Grouping them with former British colonies such as Nigeria or India would risk obscuring meaningful differences, particularly since the settler-societies built institutions oriented around greater political inclusion, public investment, and local autonomy (Acemoglu et al., 2001). We therefore exclude settler colonies from the colonial categorization altogether.

[14] Some examples of how countries bearing multiple colonial legacies were classified. India is coded as a British colony despite the Portuguese presence in Goa until 1961. Goa accounted for less than two percent of India's population, and its influence on the national civil service and university system is marginal. In contrast, British colonial rule structured India's core institutional frameworks, including its elite educational institutions and public sector hiring systems (Banerjee and Iyer, 2005). More examples are provided in Table A4 of the Appendix.

trative cases of countries with multiple colonial histories and the rationale for their classification summarized in Table A4 of the Appendix.

2.4.2 Formerly part of the Soviet Union

We identify whether a country was part of the former Soviet Union, capturing exposure to centrally planned, quantitatively oriented curricula.[15] Even after the dissolution of the USSR, many successor states retained elements of this pedagogical model, including a quantitatively intensive curriculum and institutional norms distinct from neighboring European or Asian systems. These traditions, emphasized mathematical formalism, and in our case, may predispose graduates toward fields such as econometrics or microeconomics theory. The long-term influence of Soviet mathematical frameworks on economic training is reflected in the establishment of institutions such as the Central Economic Mathematical Institute (CEMI), built to integrate mathematical methods into economic planning and economic research.

We include a binary variable indicating whether a graduate's inferred country of origin belonged to the former Union of Soviet Socialist Republics (USSR).[16]

2.4.3 Muslim Population Share

We include the share of a country's population that identifies as Muslim, to capture prevailing cultural norms around gender roles, household authority, and inter-generational expectations (Guiso et al. 2006).[17] Such norms can influence which research areas students view as accessible or appropriate, even after migration. For example, recent cross-country research shows that in societies with historically high levels of Muslim adherence, economic shocks such as natural resource booms were associated with declines in female labor force participation, while the same shocks increased women's employment in less religiously conservative settings (Joslin and Nordvik 2021). This suggests that religious composition can shape how gender norms respond to structural change, with lasting implications for women's economic roles and, by extension, their academic and professional choices. In our context, the continuous Muslim-share variable helps further disentangle these norm-driven effects from broader regional influences.

Countries with high Muslim shares span a diverse set of regions - from South and Central Asia to North Africa, the Middle East, and parts of Europe and the former Soviet Union, which the current regional classification does not capture.

[15] Some of these curricular and institutional effects likely extended beyond the USSR itself to Eastern Bloc countries more broadly (e.g.- East and West Germany (Alesina and Fuchs-Schundeln, 2007). However, we restrict our classification to former Soviet republics because the boundaries of Soviet versus domestic influence in other states might be harder to disentangle - for instance, distinguishing East from West German origin graduates or accounting for countries with mixed legacies of Soviet and national institutions. Including these cases risks conflating distinct historical trajectories, so we adopt the narrower USSR classification while recognizing its partial overlap with broader Eastern Bloc traditions.

[16] This includes the following countries: Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan.

[17] We use current Muslim population shares available in the World Population Review (2018-23, depending on availability) as proxies for social and cultural norms at origin. While contemporary, these shares are relatively stable over time in most countries and offer a reasonable approximation of the normative environments that may influence even second- or third-generation migrants.

3 Methodology

We model field specialization as the outcome of a discrete choice process in which students select the field that yields the highest expected utility given their background, training environment, and broader institutional context (similar to Sierminska and Oaxaca 2022). Formally, the utility that individual i derives from choosing field j can be written as:

$$U_{ij} = \alpha_j + \beta_{1j}\text{Region}_i + \beta_{2j}^{\top}\text{Individual}_i + \beta_{3j}^{\top}\text{Institution}_i + \beta_{4j}^{\top}\text{Country}_i + \varepsilon_{ij},$$

where U_{ij} is the latent utility of field j, α_j is a field-specific constant, and ε_{ij} is an idiosyncratic error term. Here, Region_i denotes a single categorical variable (or a set of regional indicators) capturing the student's region of origin, while Individual_i, Institution_i, and Country_i represent vectors of characteristics at the individual, institutional, and country levels, respectively. β_{1j} captures regional differences in field choice, β_{2j} reflects the effects of individual-level factors such as gender and graduation year, β_{3j} relates to institutional characteristics such as department rank, public-university status, and female-faculty share, and β_{4j} captures country-level legacies such as colonial history, former USSR membership, and religious composition.

The individual is assumed to choose the field with the highest utility, leading to a multinomial logit specification. The probability of observing student i in field j is:

$$\begin{aligned}\Pr(Y_i = j \mid X_i) &= \frac{\exp\Big(\alpha_j + \beta_{1j}\text{Region}_i + \beta_{2j}^{\top}\text{Individual}_i + \beta_{3j}^{\top}\text{Institution}_i + \beta_{4j}^{\top}\text{Country}_i\Big)}{\sum_{k=1}^{J}\exp\big(\alpha_k + \beta_{1k}\text{Region}_i + \beta_{2k}^{\top}\text{Individual}_i + \beta_{3k}^{\top}\text{Institution}_i + \beta_{4k}^{\top}\text{Country}_i\big)}, \\ &\quad \text{for } j = 1, \dots, J.\end{aligned}$$

We estimate this model by maximum likelihood and report average marginal effects, which capture the change in predicted probabilities associated with a one-unit change in an explanatory variable, averaged across the sample. Robust standard errors are clustered at the university level to account for correlated shocks within institutions.

3.1 Variables and Controls

The covariates are grouped into four broad categories:

- **Region of origin:** indicators for seven regions based on surname-inferred country of origin: North America (reference category), Europe, East Asia, South/Central Asia, Latin America and the Caribbean (LAC), Middle East/North Africa (MENA), and Sub-Saharan Africa. Regional origin proxies for students' experiences before graduate school, which may shape the salience of different economic problems (e.g., development in low- and middle-income contexts, or finance in more market-oriented systems).

- **Individual characteristics:** a female indicator variable (0/1) and year of graduation (2009–2018). Gender is included to capture systematic differences in exposure, mentorship, and perceived accessibility of different subfields. Graduation year controls for time trends in specialization patterns and ensures comparisons are not driven by changing field-level demand across cohorts.

- **Institutional environment:** indicators for department rank (top 20, 21–50, 51+), a public university dummy, and the share of female faculty in the department. These capture both the prestige of training institutions and the gender composition of the doctoral environment. Since U.S. economics PhD programs typically require students to declare their fields of specialization in the second or third year of study, some of the institutional features are likely to shape field choice through exposure to faculty role models, available course offerings, and perceptions of field-specific career prospects.
- **Country-level context:** colonial heritage (British, French, Spanish, Portuguese), former USSR membership, and the share of Muslims in the population. These variables are not treated as causal determinants, but rather as proxies for persistent institutional and cultural legacies that shape disciplinary traditions. For instance, colonial heritage often influences legal systems and educational structures, while Soviet-era training emphasized mathematical and planning-oriented fields. The religious composition potentially captures cultural norms that may affect the perceived legitimacy or relevance of particular fields of inquiry. Including these measures helps account for within-region heterogeneity when country fixed effects are infeasible given sample limitations.

3.2 Outcome Variable Definitions

Our primary outcome is a categorical indicator of field specialization, constructed from a harmonized classification of dissertation abstracts (see Section 2.3). In the main specification, each dissertation is assigned to one of ten mutually exclusive fields, prioritizing JEL codes where available and supplementing with topic-model assignments when ambiguous.

To assess robustness, we also estimate models using three alternative definitions of the dependent variable: (i) dissertations with a single JEL code; (ii) the dominant topic from the topic model; and (iii) continuous topic probabilities (theta values) from the topic model (described in Section 3.6), which capture the distribution of thematic emphasis across fields. These alternative specifications ensure results are not an artifact of classification and provide a fuller picture of field sorting.

Together, this combination of baseline, heterogeneity, and robustness specifications offers a comprehensive account of how individual, institutional, and country-of-origin characteristics shape field specialization among economics PhD graduates.

3.3 Hausman Test

A key assumption of the multinomial logit model is the Independence of Irrelevant Alternatives (IIA), which requires that the relative odds of choosing between any two outcome categories are unaffected by the presence or absence of other alternatives. Violations of IIA would imply that estimated substitution patterns across fields are biased.

To test this, we conduct a series of Hausman tests, sequentially excluding each outcome category and comparing the restricted and unrestricted estimates. In all cases, we fail to reject the null hypothesis of IIA (see Table A5), providing no statistical evidence of violation under our specification. We therefore proceed with the multinomial logit framework in all subsequent estimations.

3.4 Heterogeneity: By Gender and Region

To examine heterogeneity in field specialization patterns, we re-estimate the main multinomial logit specification separately for males and females, and then separately within each region. The graduation year variable is treated as continuous in these regional subsample models to avoid oversaturation of fixed effects and convergence issues.

For the regional subsamples, we summarize results using a compact table that records only the **signs** of statistically significant coefficients (at the 5% level), rather than reporting full estimates. This facilitates comparability across regions and highlights systematic differences without overburdening the presentation. For instance, a "+" indicates that a given predictor increases the likelihood of specialization in a field within that region, while a "–" indicates the opposite. Country-level variables are not displayed in the regional subsample tables due to limited within-region variation in some cases, though they are retained as controls in the estimation to ensure consistent specifications across all models.

3.5 Alternative Specification: Genetic Distance

As an alternative to region dummies, we estimate the main model using genetic distance to the United States as the principal explanatory variable. Genetic distance captures the historical separation of populations based on inherited traits and has been used in the literature as a proxy for cultural and institutional proximity across countries (Spolaore and Wacziarg, 2009).

The interpretation in this context is that greater genetic distance from the U.S. reflects more distinct cultural, institutional, and educational traditions, which may shape the types of economic problems viewed as salient during training. Unlike regional dummies, which impose sharp boundaries, genetic distance allows us to test whether field specialization varies along a continuous gradient of proximity to the U.S.

Formally, the model becomes:

$$Pr(Y_i = j \mid X_i) \;=\; \frac{\exp\left(\alpha_j + \gamma_j \, \text{GeneticDistance}_i + \delta_j \, Z_i\right)}{\sum_{k=1}^{J} \exp\left(\alpha_k + \gamma_k \, \text{GeneticDistance}_i + \delta_k \, Z_i\right)},$$

where GeneticDistance_i is the average genetic distance of country of origin i to the U.S., and Z_i includes the same set of individual, institutional, and country-level controls as in the baseline specification.

3.6 Alternative Outcome: Topic Modeling Probabilities

In a complementary specification, we move beyond categorical field assignments and instead treat the topic probabilities generated by the Latent Dirichlet Allocation (LDA) model as continuous outcomes. Each dissertation abstract is represented by a ten-element probability vector $\theta_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{i10})$, where θ_{ij} is the share of text associated with topic j and $\sum_{j=1}^{10} \theta_{ij} = 1$.

For example, one abstract may be 0.50 aligned with Labor, 0.25 with Trade, and 0.25 with Econometrics. The continuous θ values range from 0 to 1, and by construction their sum equals 1 for each abstract.

To exploit this richer information, we estimate a series of ordinary least squares (OLS) regressions of the form:

$$\theta_{ij} = \alpha_j + \beta_{1j}\text{Region}_i + \beta_{2j}\text{Individual}_i + \beta_{3j}\text{Institution}_i + \beta_{4j}\text{Country}_i + \varepsilon_{ij},$$

where θ_{ij} is the topic probability for field j, and the covariates mirror those in the multinomial logit model: region-of-origin indicators, gender and graduation year, institutional environment (rank, public university, female faculty share), and country-level legacies (colonial heritage, USSR membership, religious composition).

This specification enables us to assess whether particular characteristics are systematically associated with greater representation in a given topic. For instance, graduates from Sub-Saharan Africa may exhibit a higher θ value for Development and Growth topics, even if that is not their single dominant category. By retaining the full distribution of topic weights across all ten topics, this approach provides a complementary robustness check and offers a more nuanced picture of thematic specialization.

3.7 Limitations

Two caveats are worth noting in interpreting the analysis. First, we do not observe students' family backgrounds or undergraduate preparation, which limits our ability to fully account for formative influences prior to doctoral study. However, the structure of U.S. PhD programs provides a comparative advantage in studying field choice: most students undergo one to two years of shared coursework before selecting a specialization, allowing us to observe how institutional exposure, faculty interaction, and curriculum shape specialization decisions. By contrast, in many European and other systems, students must commit to a field and often a supervisor before entering the program (Cyranoski et al. 2011), making it harder to disentangle pre-existing preferences from institutional assignment. Thus, while pre-PhD factors certainly matter, the U.S.-style training model offers a cleaner window into how specialization decisions evolve during doctoral training.

Second, prior evidence suggests that drop-out rates differ by field, gender, and institutional rank (Euler et al. 2018; Stock et al. 2011), and lower-ranked programs - where students from underrepresented regions such as Sub-Saharan Africa are more concentrated - tend to have higher attrition. Thus, attrition patterns may introduce some bias. Since our sample only includes completed dissertations, we may understate certain demographic gaps that emerge earlier in the pipeline. Nevertheless, the dataset provides a rich foundation for analysis, linking dissertation content (title, abstract, JEL codes) to institutional and inferred demographic characteristics.

4 Results

4.1 Descriptive statistics

Table 6 presents summary statistics for all graduates in the main sample. Women make up 29 percent of the sample. As discussed earlier, compared to men, they are more likely to be from Europe and South Asia and less likely to be from North America. Female graduates are more likely to be in public universities and in relatively lower-ranked departments.

Table 6: Descriptive Statistics

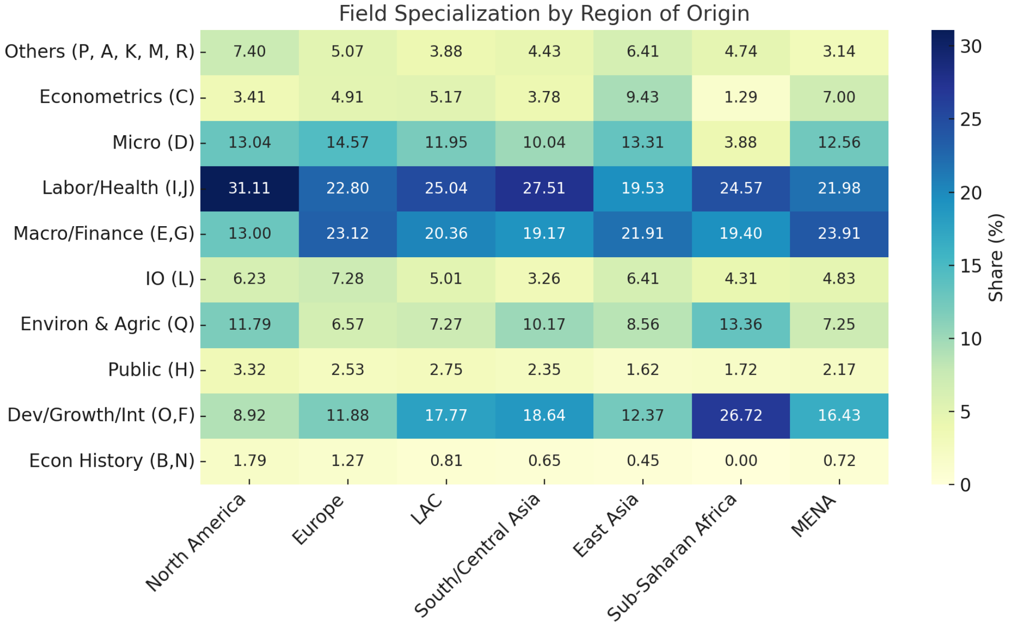
Mean / %	**Overall**	**Male**	**Female**
Female (%)	29.2	–	–
Region of Origin			
North America	27.3	29.3	22.3
Europe	15.4	14.8	16.9
LAC	7.6	8.3	5.9
South Asia / Central Asia	9.4	7.7	13.4
East Asia	32.4	32.2	32.9
Sub-Saharan Africa	2.8	2.9	2.7
MENA	5.1	4.7	5.9
University/Department and Individual Characteristics			
Rank: 1–20	39.9	41.3	36.3
Rank: 21–50	28.3	28.1	28.8
Rank: 51 and higher	31.9	30.6	34.9
Public University	57.5	56.4	60.1
Year of Graduation	2013.4	2013.5	2013.3
Share of Females in Dept.	16.6	16.4	17.0
Country-of-Origin Characteristics			
Portuguese Colony	2.5	2.6	2.2
Spanish Colony	5.2	5.8	3.7
French Colony	3.4	3.4	3.4
British Colony	12.2	10.8	15.9
Ex-USSR	3.6	3.1	5.0
Share of Muslims (%)	11.5	10.9	12.9
N	8,185	5,793	2,391

Note: For categorical variables such as region of origin, the entries represent the share of graduates within each gender group. For example, 22.3 percent of female graduates are from North America compared to 29.3 percent of males.

Figure 1 visualizes overall field specialization by region of origin, revealing interesting patterns among the most common fields. Labor/Health; Macro/Finance; Development/ Growth dominate among graduates from Sub-Saharan Africa; Macro/Finance is most popular among those from East Asia, MENA, and Europe; while Labor is the leading field for those from North America, South Asia, and LAC. East Asian graduates also exhibit a relatively high concentration in Econometrics, whereas those from Sub-Saharan Africa are less likely to specialize in

Microeconomics.

Figure 1: Field specializations by region of origin



Note: Each column shows the distribution of field specializations for graduates from a given region. Percentages within a column sum to 100, so values indicate the share of students from that region specializing in each field. For example, among all graduates from North American origin, 31 percent specialize in Labor/Health, 13 percent in Macro/Finance, and 13 percent in Micro.

Disaggregating by gender (Figure 2) adds further nuance. Labor emerges as the dominant field for women across all regions (with at least a 25% presence), though the extent of concentration differs sharply. For instance, nearly 45 percent of North American-origin women specialize in Labor, compared to just 25 percent among East Asian women, where field choices are more evenly spread. While East Asian men are more likely to enter Econometrics than women from the same region, East Asian women are still more likely to enter Econometrics than men from other regions such as South Asia, highlighting the importance of intersectional comparisons. Among men, field distributions are more diffused. No single field exceeds 30 percent in any region. Macro/Finance, however, is often dominant: it accounts for 24 percent of LAC men's specializations but just 9 percent among women from the same region. Similarly, 23 percent of East Asian men choose Macro, compared to 18.5 percent of East Asian women.

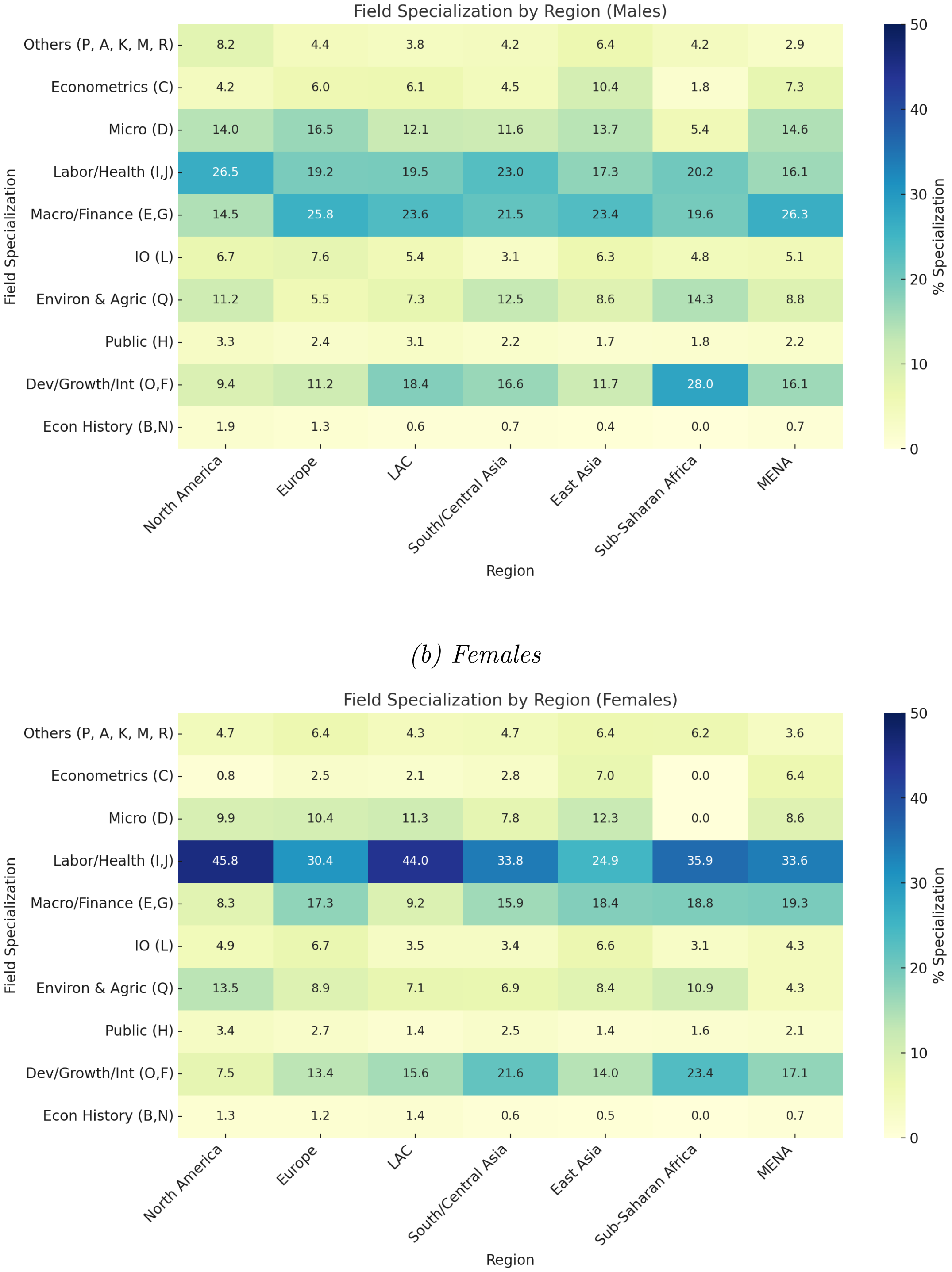
These descriptive findings already suggest two important takeaways. First, field specialization is strongly patterned by region of origin. Second, apparent gender gaps in field choice may be partly explained (and/or substantially shaped) by regional variation. We explore these relationships more systematically in the regression analysis that follows.

4.2 Regression Results

Table 7 presents results from the main multinomial regression described in Section 3. Field specialization varies systematically by region of origin. Compared to graduates from North America (the reference group), those from most other regions are significantly less likely to pursue dissertations in Labor, and more likely to specialize in Macro/Finance. The largest

Figure 2: Field specialization by region of origin and gender

(a) Males



(b) Females

Note: Each panel shows the distribution of field specializations for graduates from a given region, disaggregated by gender. Percentages within a column sum to 100, so values indicate the share of male or female students from that region specializing in each field.

difference in this regard is observed among graduates from LAC.

In Microeconomics, graduates from Sub-Saharan Africa and LAC are significantly more likely to specialize in this field relative to North American graduates. For Environment and Agriculture, African graduates are particularly overrepresented. Development, Growth, and International Economics (which we group into a single category) are especially prominent among those from Sub-Saharan Africa, East Asia, and South Asia, with Sub-Saharan Africa showing the strongest positive association in comparison to North American graduates.

At the country level, institutional legacies show consistent and significant associations with field specialization. For example, graduates with country of origin from countries formerly colonized by Portugal, Spain or Britain are significantly less likely to pursue Econometrics. By contrast, those from former Soviet countries are more likely to specialize in Econometrics and Microeconomics, and less likely to pursue fields like Labor. This aligns with the literature that links Soviet-style systems to more mathematically-oriented training. Former British colonies also show a lower likelihood of specialization in Macro/Finance and a higher likelihood in Development/Growth fields. Graduates from countries with higher share of Muslims are more likely to specialize in Macro/Finance. Most likely these are a result of both demand and supply factors.

To assess whether these results are sensitive to how field specialization is defined, we estimate the same regression using alternative outcomes. Table A6 restricts the sample to dissertations with a single self-reported JEL code, while Table A7 uses topic modeling output to classify abstracts into six broad fields aligned with the original JEL taxonomy. Robustness checks in Table A6 remain qualitatively similar: compared to graduates from North America, graduates from most regions are more likely to pursue Econometrics, and those from South Asia, East Asia, and Sub-Saharan Africa show stronger representation in Development and Growth. As in the main regression, several regions remain significantly less likely to specialize in Labor, although some coefficients lose significance due to the smaller sample size.

Table A7 confirms these broad patterns using the topic-modeling-based classification. Once again, graduates from other regions are more likely than their North American counterparts to pursue Macro/Finance and Econometrics. However, this specification shows weaker evidence of regional differences in Development and International Economics, likely reflecting differences in how development- and trade-related keywords are grouped in the topic model.

Overall, these robustness and sensitivity checks suggest that the main findings are not overly sensitive to how field specialization is defined. Whether measured using JEL codes, topic modeling, or restricted samples, region of origin characteristics remain strongly associated with the choice of field during doctoral training and show a wide variation, indicating its salience.

Table 7: Main Regression Results (Field Specializations based on JEL codes and Topic Modeling output)

	Others	**Econometrics**	**Micro**	**Labor/Health**	**Macro/Finance**	**IO**	**Environ & Agric**	**Public**	**Dev/Growth/Int**	**Econ Hist**
Ref: North America										
Europe	-0.028***	0.016*	-0.005	-0.086***	0.117***	0.010	-0.049***	-0.007	0.033**	-0.002
	(0.010)	(0.010)	(0.013)	(0.016)	(0.015)	(0.008)	(0.013)	(0.006)	(0.016)	(0.002)
LAC	-0.033*	0.088**	-0.133*	-0.088	0.177**	-0.057	0.023	0.000	0.029	-0.005*
	(0.020)	(0.043)	(0.077)	(0.110)	(0.072)	(0.051)	(0.047)	(0.009)	(0.069)	(0.003)
South & Central Asia	-0.032*	0.036**	-0.004	-0.088***	0.100***	-0.035**	-0.016	-0.015	0.062***	-0.009*
	(0.018)	(0.015)	(0.022)	(0.031)	(0.026)	(0.016)	(0.021)	(0.010)	(0.020)	(0.005)
East Asia	-0.009	0.054***	0.005	-0.143***	0.111***	0.003	-0.038***	-0.018***	0.042***	-0.007***
	(0.008)	(0.010)	(0.010)	(0.019)	(0.016)	(0.008)	(0.013)	(0.004)	(0.014)	(0.002)
Sub-Saharan Africa	-0.017	-0.013	-0.105***	-0.050	0.122***	-0.010	0.024	-0.013	0.139***	-0.077***
	(0.018)	(0.029)	(0.041)	(0.042)	(0.032)	(0.021)	(0.030)	(0.013)	(0.023)	(0.016)
MENA	-0.044**	0.055***	0.003	-0.117***	0.101***	-0.005	-0.036	-0.008	0.059**	-0.008*
	(0.022)	(0.016)	(0.029)	(0.036)	(0.030)	(0.021)	(0.022)	(0.014)	(0.028)	(0.004)
Portuguese Colony	-0.030	-0.076*	0.136*	0.028	-0.074	0.031	-0.084	0.004	0.061	0.004
	(0.020)	(0.044)	(0.076)	(0.109)	(0.067)	(0.046)	(0.058)	(0.009)	(0.067)	(0.003)
Spanish Colony	0.002	-0.075*	0.091	0.037	-0.109	0.046	-0.053	-0.008	0.068	0.001
	(0.025)	(0.045)	(0.083)	(0.120)	(0.076)	(0.053)	(0.052)	(0.012)	(0.070)	(0.004)
French Colony	0.004	-0.008	-0.042*	0.066**	-0.054**	-0.007	0.042**	-0.012	0.010	0.002
	(0.015)	(0.011)	(0.025)	(0.032)	(0.027)	(0.015)	(0.017)	(0.013)	(0.020)	(0.004)
British Colony	0.006	-0.026**	-0.022	0.027	-0.031*	0.001	0.005	0.007	0.030**	0.003
	(0.014)	(0.012)	(0.016)	(0.024)	(0.019)	(0.014)	(0.014)	(0.007)	(0.015)	(0.003)
Ex-USSR	0.022	0.019*	0.041**	-0.063*	-0.032	-0.007	0.007	0.006	0.008	-0.001
	(0.015)	(0.012)	(0.018)	(0.032)	(0.022)	(0.015)	(0.026)	(0.010)	(0.023)	(0.003)
Muslim Share	-0.016	-0.012	0.009	0.005	0.038*	-0.013	-0.018	-0.005	0.011	0.002
	(0.018)	(0.015)	(0.021)	(0.029)	(0.023)	(0.021)	(0.023)	(0.013)	(0.019)	(0.004)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table reports average marginal effects from multinomial logit regressions of field specialization on region of origin and country-level characteristics. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *All specifications include controls for department rank, share of female faculty, and whether the university is public. Standard errors are clustered at the university level.*

4.3 Topic Modeling Probabilities

In Table 8, we estimate a series of linear regressions using the topic probabilities generated from the LDA model, treating each of the ten topics as a continuous outcome ranging from 0 to 1. Unlike previous specifications that focus only on the dominant field per dissertation, this approach retains the full vector of topic probabilities and captures the extent to which each abstract emphasizes a given field. In other words, instead of measuring whether a graduate "does Labor" or "does Macro," this specification asks how much of their abstract aligns with each thematic category identified through topic modeling. This provides a more flexible view of field orientation, especially for dissertations that span multiple themes.

While this specification captures a different dimension of specialization i.e. probabilistic thematic emphasis rather than dominant field, the broad patterns reinforce earlier findings. In particular, they confirm that regional of origin differences shape not only what field students are most strongly associated with, but also the degree to which their work spans or emphasizes particular themes. Taken together, the results offer a richer picture of field orientation and its variation across origin groups.

We find areas of alignment, as well as, divergence with earlier results. Consistent with prior regressions, compared to North America, graduates from Europe, South Asia, East Asia, and MENA are more likely to have a higher share of Econometrics-related keywords in their abstracts. Topic probabilities for Finance also tend to be higher for graduates from most non-North American regions. However, separating Macroeconomics and Finance into distinct topics reveals additional variation: while LAC continues to show higher shares of Macroeconomics content, East Asia, South Asia, and Sub-Saharan Africa show significantly lower shares of Macroeconomics-related keywords - differences that were less visible in the earlier combined Macro/Finance category. It is worth noting that the topic labeled "Macroeconomics" also includes frequent references to taxation, government, and public policy, suggesting conceptual overlap with Public Economics. Hence, some of these regional differences may partly reflect variation in the prevalence of fiscal-policy or public-sector–related work rather than purely macroeconomic analysis.

Labor, Education, and Health appear as separate topics in this specification, allowing for more granular comparisons. Graduates from South Asia are more likely to emphasize Labor-related content, compared to North America, while those from East Asia show relatively lower Labor probabilities. As before, Europe, East Asia and MENA also exhibit lower shares of Education and Health content. Though, abstracts from South Asia, show significantly lower topic shares in Education. Finally, the topic identified as "Trade" (which overlaps conceptually with our earlier Development/Growth/International category) shows higher representation in East Asia and Sub-Saharan Africa, relative to North America.

Table 8: Regression Results (based on Topic Modeling probabilities)

	Agriculture	**Labor**	**Growth/Dev**	**IO/Game Theory**	**Macro**	**Health**	**Education**	**Trade**	**Finance**	**Econometrics**
Ref: North America										
Europe	-0.0265***	-0.0035	-0.0118**	0.0026	-0.0103	-0.0310***	-0.0500***	0.0151**	0.0831***	0.0323***
	(0.0073)	(0.0092)	(0.0054)	(0.0094)	(0.0091)	(0.0089)	(0.0070)	(0.0073)	(0.0108)	(0.0097)
LAC	0.0178	-0.0023	-0.0253	-0.0422	-0.0708**	0.0639	-0.0531	0.0603**	0.0220	0.0297
	(0.0285)	(0.0427)	(0.0303)	(0.0346)	(0.0327)	(0.0708)	(0.0438)	(0.0300)	(0.0918)	(0.0291)
South/Central Asia	-0.0077	0.0298*	-0.0450***	-0.0326**	-0.0153	-0.0111	-0.0453***	0.0213	0.0735***	0.0324**
	(0.0113)	(0.0158)	(0.0103)	(0.0147)	(0.0155)	(0.0220)	(0.0145)	(0.0170)	(0.0219)	(0.0141)
East Asia	-0.0222**	-0.0194**	-0.0410***	-0.0076	-0.0365***	-0.0263***	-0.0603***	0.0385***	0.0977***	0.0769***
	(0.0091)	(0.0075)	(0.0082)	(0.0085)	(0.0069)	(0.0100)	(0.0069)	(0.0075)	(0.0097)	(0.0108)
Sub-Saharan Africa	0.0523	0.0162	-0.0305**	-0.0465***	-0.0390**	-0.0246	-0.0275	0.0402**	0.0576**	0.0017
	(0.0342)	(0.0163)	(0.0123)	(0.0153)	(0.0175)	(0.0209)	(0.0178)	(0.0194)	(0.0248)	(0.0155)
MENA	-0.0133	-0.0004	-0.0338***	-0.0082	-0.0153	-0.0152	-0.0371**	0.0144	0.0634**	0.0455**
	(0.0147)	(0.0159)	(0.0101)	(0.0203)	(0.0156)	(0.0205)	(0.0169)	(0.0205)	(0.0270)	(0.0211)
	(0.0039)	(0.0055)	(0.0032)	(0.0049)	(0.0054)	(0.0060)	(0.0053)	(0.0065)	(0.0076)	(0.0063)
Portuguese Colony	-0.0654**	0.0104	0.0240	0.0154	0.0542	-0.0879	0.0223	-0.0340	0.0881	-0.0271
	(0.0278)	(0.0420)	(0.0308)	(0.0322)	(0.0378)	(0.0710)	(0.0435)	(0.0279)	(0.0986)	(0.0310)
Spanish Colony	-0.0421	0.0339	0.0171	0.0166	0.0763**	-0.0781	0.0047	-0.0355	0.0366	-0.0295
	(0.0319)	(0.0452)	(0.0303)	(0.0344)	(0.0357)	(0.0728)	(0.0428)	(0.0320)	(0.0955)	(0.0301)
French Colony	0.0066	0.0210*	0.0135*	-0.0157	0.0176	0.0279*	0.0154	-0.0075	-0.0390*	-0.0398***
	(0.0136)	(0.0117)	(0.0081)	(0.0147)	(0.0121)	(0.0142)	(0.0126)	(0.0132)	(0.0198)	(0.0144)
British Colony	-0.0128	-0.0009	0.0239***	0.0077	0.0082	0.0399***	-0.0052	0.0052	-0.0348**	-0.0313***
	(0.0077)	(0.0118)	(0.0079)	(0.0115)	(0.0118)	(0.0125)	(0.0089)	(0.0120)	(0.0157)	(0.0107)
Ex-USSR	-0.0071	-0.0093	-0.0062	0.0251	-0.0057	-0.0111	-0.0102	0.0232*	-0.0327*	0.0340**
	(0.0123)	(0.0168)	(0.0078)	(0.0191)	(0.0124)	(0.0139)	(0.0106)	(0.0137)	(0.0186)	(0.0160)
Muslim Share	-0.0301**	-0.0164	0.0022	-0.0082	0.0063	-0.0214	-0.0052	0.0246	0.0518**	-0.0037
	(0.0122)	(0.0137)	(0.0114)	(0.0143)	(0.0148)	(0.0160)	(0.0126)	(0.0177)	(0.0205)	(0.0169)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Other controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	6127	6127	6127	6127	6127	6127	6127	6127	6127	6127

Note: Each outcome represents a probability (0–1) of the dissertation abstract belonging to a given field. An abstract may span multiple fields. All specifications include controls for department rank, share of female faculty, whether the university is public, and country-level characteristics. Standard errors are clustered at the university level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.4 Gender

In Tables 9 and 10, we present the results of the main regression (from Table 7) estimated separately for males and females. While several regional patterns remain consistent across genders, important divergences also emerge.

For instance, female graduates of European origin are significantly more likely to specialize in Econometrics compared to their North American-origin counterparts, a difference that is not statistically significant among males. Similarly, while men from LAC are more likely to pursue Macro/Finance relative to U.S. males, this pattern does not hold for women. In contrast, women from LAC, Europe, and MENA are significantly more likely to specialize in Development, Growth, and International Economics compared to U.S.-origin females, a difference not observed among males. The patterns observed in Labor Economics appear to be broadly consistent across genders: graduates from most non-U.S. regions are less likely to specialize in Labor compared to U.S. graduates. However, one exception stands out - women of Sub-Saharan African origin are more likely to pursue Labor, marking a notable deviation from the general pattern of underrepresentation in this field.

When considering country-level variables, we find further gender-based variation. Among male graduates, coming from a country formerly part of the USSR is associated with a greater likelihood of specializing in more technical fields such as Econometrics and Microeconomics. This association is not statistically significant for women, suggesting that the influence of prior Soviet institutional legacy on field choice may operate more strongly among men, perhaps reflecting differential selection, educational exposure, or incentives linked to these fields.

Overall, the gender-disaggregated results highlight how both gender and regional origin intersect in shaping field specialization. These differences underscore the importance of examining field choices through an intersectional lens, rather than attributing variation to either gender or origin alone.

Table 9: Main Regression Results (Males only)

	Others	**Econometrics**	**Micro**	**Labor/Health**	**Macro/Finance**	**IO**	**Environ & Agric**	**Public**	**Dev/Growth/Int**	**Econ Hist**
Ref: North America										
Europe	-0.035***	0.014	-0.002	-0.062***	0.121***	0.011	-0.065***	-0.007	0.026	-0.001
	(0.012)	(0.013)	(0.015)	(0.019)	(0.018)	(0.011)	(0.017)	(0.007)	(0.017)	(0.002)
LAC	-0.029	0.113*	-0.165*	-0.096	0.204**	-0.078	0.058	0.003	-0.005	-0.005
	(0.024)	(0.058)	(0.088)	(0.126)	(0.095)	(0.061)	(0.058)	(0.012)	(0.068)	(0.004)
South/Central Asia	-0.035	0.038*	-0.016	-0.070**	0.089***	-0.024	-0.004	-0.022*	0.053**	-0.008**
	(0.023)	(0.020)	(0.025)	(0.033)	(0.033)	(0.021)	(0.024)	(0.012)	(0.025)	(0.004)
East Asia	-0.015	0.055***	-0.005	-0.107***	0.104***	-0.002	-0.034**	-0.018***	0.029**	-0.007***
	(0.009)	(0.011)	(0.012)	(0.020)	(0.018)	(0.009)	(0.014)	(0.005)	(0.014)	(0.002)
Sub-Saharan Africa	-0.025	-0.007	-0.098**	-0.055	0.106**	0.002	0.025	-0.017	0.137***	-0.067***
	(0.021)	(0.036)	(0.044)	(0.048)	(0.043)	(0.021)	(0.032)	(0.014)	(0.024)	(0.015)
MENA	-0.059**	0.061***	0.013	-0.123***	0.091**	0.012	-0.014	-0.014	0.037	-0.005
	(0.025)	(0.019)	(0.035)	(0.047)	(0.037)	(0.022)	(0.023)	(0.017)	(0.029)	(0.003)
Portuguese Colony	-0.050**	-0.101*	0.172*	0.004	-0.072	0.041	-0.099	0.000	0.102	0.001
	(0.024)	(0.059)	(0.092)	(0.119)	(0.090)	(0.055)	(0.067)	(0.010)	(0.064)	(0.004)
Spanish Colony	-0.008	-0.102*	0.109	0.043	-0.121	0.073	-0.085	-0.006	0.097	0.000
	(0.029)	(0.059)	(0.092)	(0.133)	(0.099)	(0.062)	(0.062)	(0.015)	(0.068)	(0.005)
French Colony	-0.022	-0.006	-0.028	0.061*	-0.031	0.003	0.049**	-0.005	-0.021	-0.001
	(0.023)	(0.015)	(0.027)	(0.035)	(0.030)	(0.017)	(0.020)	(0.013)	(0.023)	(0.005)
British Colony	-0.006	-0.030*	-0.001	0.043*	-0.025	-0.015	0.015	0.010	0.005	0.004
	(0.017)	(0.019)	(0.021)	(0.026)	(0.022)	(0.017)	(0.016)	(0.008)	(0.018)	(0.003)
Ex-USSR	0.000	0.035**	0.052**	-0.077*	-0.038	-0.008	0.040	-0.001	0.000	-0.004
	(0.021)	(0.016)	(0.023)	(0.044)	(0.029)	(0.021)	(0.032)	(0.013)	(0.027)	(0.004)
Muslim Share	-0.005	-0.031	-0.001	0.007	0.051*	-0.034	-0.020	0.003	0.032	-0.002
	(0.019)	(0.020)	(0.026)	(0.034)	(0.029)	(0.025)	(0.025)	(0.014)	(0.022)	(0.006)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table reports average marginal effects from multinomial logit regressions of field specialization on region of origin and country-level characteristics. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *All specifications include controls for department rank, share of female faculty, and whether the university is public. Standard errors are clustered at the university level.*

Table 10: Main Regression Results (Females only)

	Others	**Econometrics**	**Micro**	**Labor/Health**	**Macro/Finance**	**IO**	**Environ & Agric**	**Public**	**Dev/Growth/Int**	**Econ Hist**
Ref: North America										
Europe	-0.002	0.024*	-0.010	-0.172***	0.113***	0.011	-0.027*	-0.003	0.066**	0.000
	(0.016)	(0.013)	(0.015)	(0.035)	(0.028)	(0.017)	(0.016)	(0.003)	(0.034)	(0.001)
LAC	-0.030	0.013	-0.003	-0.137	0.031	-0.010	-0.091*	-0.001	0.228*	-0.001
	(0.037)	(0.020)	(0.048)	(0.255)	(0.087)	(0.035)	(0.052)	(0.008)	(0.127)	(0.001)
South/Central Asia	-0.026	0.033***	0.020	-0.150***	0.128***	-0.052**	-0.029	-0.002	0.079**	-0.002
	(0.029)	(0.012)	(0.033)	(0.057)	(0.045)	(0.026)	(0.042)	(0.007)	(0.038)	(0.002)
East Asia	0.011	0.046***	0.016	-0.255***	0.135***	0.017	-0.049**	-0.007**	0.086***	-0.001*
	(0.014)	(0.010)	(0.012)	(0.031)	(0.026)	(0.013)	(0.019)	(0.003)	(0.028)	(0.001)
Sub-Saharan Africa	0.078**	-0.262***	-1.003***	0.370***	0.351***	0.033	0.127**	0.007	0.316***	-0.016***
	(0.033)	(0.051)	(0.094)	(0.080)	(0.048)	(0.041)	(0.055)	(0.011)	(0.055)	(0.006)
MENA	-0.015	0.040***	-0.008	-0.150**	0.122***	-0.035	-0.068	0.000	0.116**	-0.002
	(0.036)	(0.015)	(0.033)	(0.066)	(0.044)	(0.037)	(0.053)	(0.008)	(0.051)	(0.002)
Portuguese Colony	0.015	0.005	0.000	0.149	-0.002	0.016	-0.021	0.003	-0.166	0.002
	(0.037)	(0.018)	(0.048)	(0.265)	(0.083)	(0.036)	(0.056)	(0.008)	(0.136)	(0.001)
Spanish Colony	0.030	0.011	0.007	0.145	-0.013	-0.030	0.069	-0.126***	-0.094	0.001
	(0.048)	(0.018)	(0.056)	(0.266)	(0.097)	(0.051)	(0.060)	(0.023)	(0.134)	(0.002)
French Colony	0.046**	-0.005	-0.047	0.131**	-0.102*	-0.036	0.048*	-0.122***	0.087**	0.001
	(0.023)	(0.011)	(0.032)	(0.058)	(0.059)	(0.040)	(0.029)	(0.023)	(0.039)	(0.002)
British Colony	0.030	-0.010	-0.048**	-0.006	-0.047	0.026	-0.024	0.000	0.079***	0.000
	(0.021)	(0.008)	(0.021)	(0.046)	(0.036)	(0.022)	(0.035)	(0.007)	(0.027)	(0.001)
Ex-USSR	0.045**	-0.003	0.020	-0.030	-0.011	-0.001	-0.050	0.005	0.025	0.001
	(0.018)	(0.013)	(0.021)	(0.061)	(0.036)	(0.026)	(0.041)	(0.005)	(0.044)	(0.001)
Muslim Share	-0.027	0.010	0.015	0.023	0.030	0.027	-0.050	-0.006	-0.024	0.002
	(0.032)	(0.010)	(0.026)	(0.061)	(0.033)	(0.030)	(0.048)	(0.009)	(0.038)	(0.001)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table reports average marginal effects from multinomial logit regressions of field specialization on region of origin and country-level characteristics. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *All specifications include controls for department rank, share of female faculty, and whether the university is public. Standard errors are clustered at the university level.*

4.5 Genetic Distance

In Table 11, we replace the region of origin categorization with a continuous measure of genetic distance to the United States, as described in Section 2.2. This specification allows us to test whether the probability of field specialization varies systematically along a gradient of long-run institutional and cultural proximity, rather than across discrete regional categories. The results reveal that a greater genetic distance from the U.S., is associated with a higher likelihood of specialization in fields such as Econometrics, Macro/Finance, and Development/Growth/International Economics. Conversely, individuals from countries genetically closer to the U.S. are more likely to specialize in Labor, Public Economics, and Economic History.

These findings broadly complement the region-based results while offering a more structured, continuous perspective. The fact that specific fields (e.g., Econometrics, Macro, Development) are positively associated with greater distance from U.S. ancestral roots suggests that students from more culturally and institutionally distant contexts may self-select (or be channeled) into certain fields based on prior training, norms, or perceived accessibility. Meanwhile, closer proximity to the U.S. appears to align with higher representation in fields such as Labor or Public Economics.

Table 11: Regression Results (Genetic Distance to the US)

	Econometrics	**Micro**	**Labor/ Health**	**Macro/ Finance**	**IO**	**Environ/ Agric**	**Public**	**Dev/ Growth/ Int**	**Econ Hist**
Genetic Distance to US	1.214*** (0.231)	-0.004 (0.224)	-3.067*** (0.446)	2.114*** (0.345)	0.035 (0.166)	-0.596* (0.310)	-0.374*** (0.104)	1.063*** (0.296)	-0.265*** (0.063)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The field specialization outcomes are from the main specification (based on a combination of JEL codes and Topic Modeling output) given in Table 7. Here, region of origin is replaced with the genetic distance variable (computed using the 'distance' of the graduates' country of origin to the US). The table reports average marginal effects from multinomial logit regressions. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All specifications include controls for department rank, share of female faculty, whether the university is public, and country-level characteristics. Standard errors are clustered at the university level.

4.6 Region-wise Regressions

In Table 12, we estimate the main regression separately by region of origin and summarize the direction and significance of marginal effects across key predictors. This disaggregated analysis highlights the considerable heterogeneity in how individual and institutional characteristics relate to field specialization across regions. Gender is a salient predictor, but its influence varies by context. In all regions, women are more likely to specialize in Labor. Being female is negatively associated with specializing in Econometrics for those from North America, Europe, East Asia, and Sub-Saharan Africa, while for Microeconomics, the negative association holds primarily for North America and Europe. The share of females in the doctoral department also interacts differently across regions and fields. A higher female share is positively associated with specializing in Labor and Development/Growth among graduates of Sub-Saharan African origin, while for those from MENA, it is positively linked with Economic History. Conversely, in North America, a higher female share is negatively associated with pursuing Economic History, and in East Asia, with pursuing Industrial Organization (IO).

University rank and type show further variation. Attending a lower-ranked institution is associated with a greater likelihood of specializing in Development/Growth among East Asian graduates, in Industrial Organization for those from the MENA, and in Microeconomics for LAC/South American graduates. Lower rank is also positively associated with pursuing Labor and Health-related fields across multiple regions. Meanwhile, attending a public university is negatively associated with specialization in Micro and Macro/Finance in several regions, but positively associated with Labor and Environmental Economics for Europeans, and with Econometrics among Sub-Saharan African graduates.

Historical and religious factors add further nuance. Originating from a former Soviet Union country is positively associated with specialization in Microeconomics and Econometrics, and negatively with Labor for those from European origin, and positively associated with Development/Growth and Environmental/Agricultural Economics among those from South and Central Asia. Similarly, a higher Muslim population share in the country of origin is positively associated with specializing in Labor for Europeans, and with Public Economics and Development/Growth for those of East Asian origin.

Taken together, these results highlight how the effects of individual identity and institutional context are not uniform but are mediated by region of origin. They indicate the importance of examining these predictors in interaction with broader cultural and institutional backgrounds, and suggest that field specialization is shaped by intersecting factors - of gender, origin, and academic environment.

Table 12: Summary of Region-wise Regression Results of Field Choice

	Female	**Share Female in Dept.**	**Rank 20–50**	**Rank 51+**	**Public Univ.**	**Year**	**Ex-USSR**	**Muslim Share**
US & Canada								
Others	-					-		
Econometrics (C)	-		-					+
Micro (D)	-		-					+
Labor/Health (I,J)	+					+		
Macro/Finance (E,G)	-							
IO (L)			+		-			-
Environ & Agric (Q)								+
Public(H)								
Dev/Growth/Int (O,F)								-
Econ History (B,N)		-						
Europe								
Others							+	
Econometrics (C)	-			-			+	
Micro (D)	-				-		+	
Labor/Health (I,J)	+				+	+	-	+
Macro/Finance (E,G)	-				-			
IO (L)								
Environ & Agric (Q)	+				+			
Public(H)								
Dev/Growth/Int (O,F)								
Econ History (B,N)								
Latin America/ South America (LAC)								
Others								
Econometrics (C)								
Micro (D)				+				
Labor/Health (I,J)	+							
Macro/Finance (E,G)	-							
IO (L)								
Environ & Agric (Q)								
Public(H)						-		-
Dev/Growth/Int (O,F)		+	-					+
Econ History (B,N)						-		-
South Asia / Central Asia								
Others							-	
Econometrics (C)								
Micro (D)				-				
Labor/Health (I,J)	+							
Macro/Finance (E,G)					-			+
IO (L)								
Environ & Agric (Q)	-						+	
Public(H)								
Dev/Growth/Int (O,F)							+	
Econ History (B,N)							-	
East Asia								
Others						+		
Econometrics (C)	-			-		-		-
Micro (D)				-	-			
Labor/Health (I,J)	+							
Macro/Finance (E,G)	-							
IO (L)		-				-		
Environ & Agric (Q)								
Public(H)								
Dev/Growth/Int (O,F)				+				+
Econ History (B,N)						+		+
Sub-Saharan Africa								
Others								
Econometrics (C)	-				+	-		
Micro (D)		-						
Labor/Health (I,J)	+	+	-			+		
Macro/Finance (E,G)					-			
IO (L)								
Environ & Agric (Q)			+					
Public(H)								
Dev/Growth/Int (O,F)		+				-		
Econ History (B,N)								
Middle East, West Asia, and North Africa (MENA)								
Others								
Econometrics (C)								
Micro (D)				-				
Labor/Health (I,J)	+		+			+		
Macro/Finance (E,G)			-					+
IO (L)			+	+		-		
Environ & Agric (Q)	-							-
Public(H)								
Dev/Growth/Int (O,F)								
Econ History (B,N)		+						

Note: Each cell reports the sign of the average marginal effect from region-specific multinomial logit regressions, estimated following the specification in the methodology and the main regression table. Only coefficients statistically significant at the 5% level are shown. For brevity, only selected fields with at least one significant coefficient are shown and we also exclude the Portuguese colonial group from regressions due to convergence issues arising from small sample sizes. "+" indicates a positive association, "–" a negative association, and blank cells indicate no significant effect.

5 Discussion and Conclusions

In this paper, we provide novel evidence on origin and identity in shaping field specialization choices for early-career economics PhD students in the United States. Rather than emerging solely from individual preferences, field choices correlate strongly with region and country of origin, institutional context, and gender. Across specifications, we find that students from outside North America are less likely to specialize in Labor Economics and more likely to concentrate in fields such as Econometrics, Macro/Finance, and Development/Growth. These patterns remain visible across multiple field classification approaches and are reinforced when we use continuous measures of genetic distance.

Some of the sharpest differences appear across regional groups. Students from Latin America are disproportionately represented in Macro/Finance, while those from Sub-Saharan Africa, South Asia, and East Asia are more likely to specialize in Development/Growth. Labor Economics, by contrast, is far more common among North American-origin students than among their peers from other regions. Country-level legacies further illuminate these associations. Former Soviet countries are more strongly linked to quantitatively intensive fields, echoing the historically mathematical orientation of Soviet economics curricula. Former British colonies show greater representation in Development/Growth fields, while higher Muslim population shares correlate with greater representation in Macro/Finance. Although sometimes the mechanisms behind these associations are not always clear - such as the link between Muslim share and Macroeconomics fields - they point to the need for further research on how inherited curricula, norms, and perceived field prestige shape specialization decisions, not only looking at supply, as is the case here, but also demand factors.

These results also extend the literature on gender and specialization. Prior work has shown that women are underrepresented in Macroeconomics and technical subfields and overrepresented in applied areas (Sierminska and Oaxaca 2021, 2022; Chari and Goldsmith-Pinkham 2022). Our analysis confirms these patterns but shows they are far from uniform. Gender differences vary substantially by region of origin: for example, Sub-Saharan African women are more likely to specialize in Labor, than those from North America, in contrast to women from most other regions; European women (but not men) are disproportionately represented in Econometrics; and the quantitative tilt associated with former Soviet origins is evident among men but not women. These examples illustrate the importance of examining gender not in isolation but in interaction with regional and institutional background. Put differently, some of the gender gaps documented in earlier studies may in part reflect differences in regional composition.

Taken together, three contributions emerge. First, region and country of origin are strongly associated with field specialization, pointing to the importance of inherited institutional and cultural factors in shaping doctoral training. Second, gender differences in specialization vary systematically by region, revealing new intersectional patterns that refine our understanding of how women and men navigate the economics discipline. Third, country-level legacies - including colonial history, religious composition, and Soviet influence - are also linked to systematic differences in field choice, suggesting that field sorting reflects multiple overlapping influences rather than uniform preferences.

Although our data cannot isolate causal mechanisms, the persistence of these patterns across classification schemes and robustness checks indicates that they are not just the result of how fields and countries are classified. Instead, they likely reflect a combination of prior exposure, disciplinary norms, training, and the signaling value of certain fields in the U.S. context. These findings also carry practical implications. Field specialization has long-term effects on placement, publication opportunities, and the kinds of policy questions that receive attention. Supporting underrepresented students therefore requires interventions that account for heterogeneous pathways into fields. Efforts to increase participation in underrepresented subfields - such as Macro/Finance or Econometrics - should recognize that gaps are larger for some groups than others. More tailored approaches could include mentoring programs that match students with field-specific role models, greater diversity in faculty recruitment across subfields, and advising practices that acknowledge diverse scholarly trajectories. While there is no "optimal" distribution of field choices, understanding why certain groups cluster in specific fields is critical for improving equity and for broadening the intellectual scope of the economics profession.

6 References

References

[1] Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5 (December 2001): 1369–1401. https://doi.org/10.1257/aer.91.5.1369.

[2] Advani, Arun, Elliott Ash, David Cai, and Imran Rasul. "Economics and the Study of Race." *VoxEU.org* (blog), May 25, 2021. https://voxeu.org/article/economics-and-study-race.

[3] Advani, Arun, Sonkurt Sen, and Ross Warwick. "Ethnic Diversity in UK Economics." October 26, 2020. https://doi.org/10.1920/BN.IFS.2020.BN0307.

[4] Alesina, Alberto, and Nicola Fuchs-Schündeln. "Good-bye Lenin (or not?): The Effect of Communism on People's Preferences." *American Economic Review* 97, no. 4 (2007): 1507–1528.

[5] Ambrosino, Angela, Mario Cedrini, John B. Davis, Stefano Fiori, Marco Guerzoni, and Massimiliano Nuccio. "What Topic Modeling Could Reveal about the Evolution of Economics." *Journal of Economic Methodology* 25, no. 4 (October 2018): 329–348. https://doi.org/10.1080/1350178X.2018.1529215.

[6] Antman, Francisca, Kirk Doran, Xuechao Qian, and Bruce A. Weinberg. "Demographic Diversity and Economic Research: Fields of Specialization and Research on Race, Ethnicity, and Inequality." SSRN Scholarly Paper, May 1, 2024. https://papers.ssrn.com/abstract=4826044.

[7] Antman, Francisca, Brian Duncan, and Stephen J. Trejo. "Ethnic Attrition and the Observed Health of Later-Generation Mexican Americans." SSRN Scholarly Paper, 2025. https://papers.ssrn.com/abstract=2819343.

[8] Banerjee, Abhijit, and Lakshmi Iyer. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review* 95, no. 4 (September 2005): 1190–1213. https://doi.org/10.1257/0002828054825574.

[9] Beneito, Pilar, José E. Boscá, Javier Ferri, and Manu García. "Gender Imbalance across Subfields in Economics: When Does It Start?" *Journal of Human Capital* 15, no. 3 (September 2021): 469–511. https://doi.org/10.1086/715581.

[10] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (March 2003): 993–1022.

[11] Chari, Anusha, and Paul Goldsmith-Pinkham. "Gender Representation in Economics Across Topics and Time: Evidence from the NBER Summer Institute." Working Paper 23953. National Bureau of Economic Research, 2022. https://doi.org/10.3386/w23953.

[12] Cyranoski, David, Natasha Gilbert, Heidi Ledford, Anjali Nayar, and Mohammed Yahia. “Education: The PhD Factory.” *Nature* 472, no. 7343 (April 2011): 276–279. https://doi.org/10.1038/472276a.

[13] Ferrara, Alessandro, and Renee Luthra. “Explaining the Attainment of the Second-Generation: When Does Parental Relative Education Matter?” *Social Science Research* 120 (2024): 103016.

[14] Fortin, Nicole, Thomas Lemieux, and Marit Rehavi. “Gender Differences in Fields of Specialization and Placement Outcomes among PhD in Economics.” American Economic Association, 2021.

[15] Freeman, Richard B., and Wei Huang. “Collaborating with People Like Me: Ethnic Coauthorship within the United States.” *Journal of Labor Economics* 33, no. S1 (July 2015): S289–318. https://doi.org/10.1086/678973.

[16] Freeman, Richard B., Danxia Xie, Hanzhe Zhang, and Hanzhang Zhou. “High and Rising Institutional Concentration of Award-Winning Economists.” 2024. https://hanzhezhang.github.io/research/2407EconAwardConcentration.pdf.

[17] Guiso, Luigi, Paola Sapienza, and Luigi Zingales. “Does Culture Affect Economic Outcomes?” *Journal of Economic Perspectives* 20, no. 2 (June 2006): 23–48. https://doi.org/10.1257/jep.20.2.23.

[18] Hofstra, Bas, Daniel A. McFarland, Sanne Smith, and David Jurgens. “Diversifying the Professoriate.” *Socius* 8 (2022): 23780231221085118. https://doi.org/10.1177/23780231221085118.

[19] Jayachandran, Seema. “Social Norms as a Barrier to Women’s Employment in Developing Countries.” *IMF Economic Review* 69, no. 3 (September 2021): 576–595. https://doi.org/10.1057/s41308-021-00140-w.

[20] Joslin, Knut-Eric, and Frode Martin Nordvik. “Does Religion Curtail Women during Booms? Evidence from Resource Discoveries.” *Journal of Economic Behavior & Organization* 187 (July 2021): 205–224. https://doi.org/10.1016/j.jebo.2021.04.026.

[21] Kerr, Sari Pekkala, and William R. Kerr. “Global Collaborative Patents.” *Economic Journal* 128, no. 612 (2018): F235–72. https://doi.org/10.1111/ecoj.12369.

[22] Kozlowski, Diego, Vincent Larivière, Cassidy R. Sugimoto, and Thema Monroe-White. “Intersectional Inequalities in Science.” *Proceedings of the National Academy of Sciences* 119, no. 2 (2022a): e2113067119. https://doi.org/10.1073/pnas.2113067119.

[23] Kozlowski, Diego, Dakota S. Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R. Sugimoto. “Avoiding Bias When Inferring Race Using Name-Based Approaches.” *PLOS ONE* 17, no. 3 (2022b): e0264270. https://doi.org/10.1371/journal.pone.0264270.

[24] Lockhart, Jeffrey W., Molly M. King, and Christin Munsch. "Name-Based Demographic Inference and the Unequal Distribution of Misrecognition." *Nature Human Behaviour* 7, no. 7 (2023): 1084–1095. https://doi.org/10.1038/s41562-023-01587-9.

[25] Markevich, Andrei, and Mark Harrison. "Great War, Civil War, and Recovery: Russia's National Income, 1913 to 1928." *Journal of Economic History* 71, no. 3 (September 2011): 672–703. https://doi.org/10.1017/S0022050711001884.

[26] May, Ann Mari, Mary G. McGarvey, and Robert Whaples. "Are Disagreements among Male and Female Economists Marginal at Best?: A Survey of AEA Members and Their Views on Economics and Economic Policy." *Contemporary Economic Policy* 32, no. 1 (January 2014): 111–132. https://doi.org/10.1111/coep.12004.

[27] Mester, Loretta J. "Increasing Diversity, Inclusion, and Opportunity in Economics: Perspectives of a Brown-Eyed Economist." Speech at the Second Annual Women in Economics Symposium, Federal Reserve Bank of St. Louis, February 28, 2019. https://ideas.repec.org/p/fip/fedcsp/107.html.

[28] Onder, Ali, and Hakan Yilmazkuday. "Thirty-Five Years of Peer-Reviewed Publishing by North American Economics PhDs: Quantity, Quality, and Beyond." *Open Economics* 3, no. 1 (2020): 70–85.

[29] Sierminska, Eva, and Ronald L. Oaxaca. "Field Specializations among Beginning Economists: Are There Gender Differences?" *AEA Papers and Proceedings* 111 (May 2021): 86–91. https://doi.org/10.1257/pandp.20211030.

[30] Sierminska, Eva, and Ronald L. Oaxaca. "Gender Differences in Economics PhD Field Specializations with Correlated Choices." *Labour Economics* 79 (December 2022): 102289. https://doi.org/10.1016/j.labeco.2022.102289.

[31] Singhal, Karan, and Eva Sierminska "Inequality in Economics as a Profession." In: Zimmermann, K.F. (eds) *Handbook of Labor, Human Resources and Population Economics* https://doi.org/10.1007/978-3-319-57365-6_453-1

[32] Sokoloff, Kenneth L., and Stanley L. Engerman. "Institutions, Factor Endowments, and Paths of Development in the New World." *Journal of Economic Perspectives* 14, no. 3 (September 2000): 217–232. https://doi.org/10.1257/jep.14.3.217.

[33] Spolaore, Enrico, and Romain Wacziarg. "The Diffusion of Development." *Quarterly Journal of Economics* 124, no. 2 (May 2009): 469–529. https://doi.org/10.1162/qjec.2009.124.2.469.

[34] Spolaore, Enrico, and Romain Wacziarg. "Ancestry, Language and Culture." Working Paper 21242. National Bureau of Economic Research, June 2015. https://doi.org/10.3386/w21242.

[35] Stansbury, Anna. "Economics Needs More Socioeconomic Diversity." *Harvard Business Review*, June 2022. https://hbr.org/2022/06/economics-needs-more-socioeconomic-diversity.

[36] Stansbury, Anna, and Robert Schultz. “The Economics Profession’s Socioeconomic Diversity Problem.” *Journal of Economic Perspectives* 37, no. 4 (December 2023): 207–230. https://doi.org/10.1257/jep.37.4.207.

[37] Stock, Wendy A., John J. Siegfried, T. Aldrich Finegan, David Colander, N. Gregory Mankiw, Melissa P. McInerney, and James M. Poterba. “Completion Rates and Time-to-Degree in Economics PhD Programs [with Comments].” *American Economic Review* 101, no. 3 (2011): 176–193.

[38] Witteman, Holly O., Jenna Haverfield, and Cara Tannenbaum. “COVID-19 Gender Policy Changes Support Female Scientists and Improve Research Quality.” *Proceedings of the National Academy of Sciences* 118, no. 6 (February 2021): e2023476118. https://doi.org/10.1073/pnas.2023476118.

[39] Zafar, Basit. “College Major Choice and the Gender Gap.” *Journal of Human Resources* 48, no. 3 (2013): 545–595.

Funding:
We gratefully acknowledge seed-grant support from the Centre of Competence for Data Science and Simulation (DSS), LISER.

Acknowledgments

We acknowledge helpful comments from participants at the University of Luxembourg Jamboree, the LISER Working Group, the AEA/ASSA Annual Meeting, the Association for Social Economics Summer School, the LIS Drivers of Inequality Conference, and seminar participants at FLAME University (India). We are also grateful to Conchita D'Ambrosio, Upasak Das, Jongoh Kim, Silvia Peracchi, Thiago Quaresma Brant Chaves, Gautam Tripathi, Philippe Van Kerm, Alexander Yarkin, and Skerdilajda Zanaj for insightful suggestions. All remaining errors are our own.

Appendix

Note A1. Country of Origin Assignment

We infer country of origin using surname data, beginning with *Forebears.io*, a global genealogy database that reports the relative frequency and density of each surname across countries. For each graduate, we assign the country in which the surname is most prevalent, provided that this top country accounts for at least 60 percent of all global occurrences. For names predicted to originate from the United States, the 80 percent threshold was necessary to avoid misclassification of surnames such as "Murphy," "Giordano," or "Smith," which are common in both American and European contexts. A similar high-threshold approach was adopted for Hispanic surnames, which are prevalent in both Iberia and Latin America.

If the above thresholds are not met, we turn to secondary evidence, examined in an order of precedence. This includes: (1) explicit statements of nationality or place of birth on curriculum vitae; (2) declarations of native language in professional biographies; (3) country of undergraduate education; (4) citizenship listings, when they differ from U.S. citizenship; and (5) affiliations with national economists' associations or other relevant indicators.

If none of these sources provide conclusive information and Forebears probabilities remain ambiguous, we deploy a large-language-model prompt using ChatGPT with search enabled. The prompt takes the form: `''Identify the most likely country of origin for economist <Name> who completed a PhD from <University>.''` If the model's result matches with the dominant category observed in Forebears, the case is resolved to reflect the dominant prediction of the algorithm.

To validate the overall classification procedure, we conducted a manual audit of 1,000 randomly selected CVs, comparing algorithmic predictions with stated origin, native language, citizenship, and undergraduate location. Based on manual insights, matches were also considered reliable when the top country exceeded 40 percent with a margin of at least ten percentage points over the next highest-ranked country.

In cases where the top three predicted Forebears countries belonged to the same world region and their combined share exceeded 60 percent, the highest-probability country was selected without further checks, since the incidence of these cases is small and regional classification - rather than exact country of origin - is the relevant analytic variable.

Further exceptions were made to names predicted to originate from the U.S. For example, if "Murphy" returned Forebears frequencies of 55 percent for the United States, 30 percent for Ireland, and 10 percent for the United Kingdom, but the CV listed undergraduate education at University College Dublin, the country of origin was assigned as Ireland. In cases of dual citizenship, such as Indian-American, we assign the non-U.S. country to preserve the idea of external cultural and institutional transmission. Where conflicting information arises, we use the order of precedence described above.

We acknowledge the possibility of misclassification using this approach, particularly among individuals with multiple ethnic backgrounds or from highly assimilated minorities. These issues and their implications are discussed further in Section 2.2 of the main text.

Table A1: Country of Origin and Region Assignment

Region of Origin	**Countries**
North America	Canada, United States, Australia, New Zealand
Europe	Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, England, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Kosovo, Latvia, Lithuania, Luxembourg, Moldova, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russia, Scotland, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Transnistria, Ukraine, Wales
Latin America & Caribbean (LAC)	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Paraguay, Peru, Trinidad and Tobago, Uruguay, Venezuela
South Asia / Central Asia	Afghanistan, Bangladesh, India, Iran, Kazakhstan, Kyrgyzstan, Nepal, Pakistan, Sri Lanka, Tajikistan, Turkmenistan, Uzbekistan
East Asia	Cambodia, China, Hong Kong, Indonesia, Japan, Laos, Macau, Malaysia, Mongolia, Myanmar, North Korea, Philippines, Singapore, South Korea, Taiwan, Thailand, Vietnam
Sub-Saharan Africa	Angola, Benin, Burkina Faso, Cameroon, Cape Verde, DR Congo, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Ivory Coast, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Somalia, South Africa, Tanzania, Togo, Uganda, Zambia, Zimbabwe
Middle East, N. Africa & E. Asia (MENA)	Algeria, Armenia, Azerbaijan, Cyprus, Egypt, Georgia, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Turkey, Yemen

Table A2: Comparing Topic Modeling Output with JEL Codes (Single Fields)

	JEL Codes									
	Econometrics	**Micro**	**Labor/ Health**	**Macro/ Finance**	**IO**	**Environ & Agri**	**Public**	**Dev/ Growth**	**Econ Hist**	**Others**
Agriculture	1.08	2.16	0.89	0.53	16.83	61.92	2.05	2.03	2.13	4.56
Development/Growth/Trade	1.43	5.98	2.03	4.10	19.03	19.64	6.16	47.98	51.08	10.54
Econometrics	79.93	25.25	3.90	7.79	5.71	3.01	4.11	1.35	2.13	6.27
Finance	5.73	8.47	1.14	72.84	1.90	5.21	8.90	20.41	14.89	12.82
IO / Game theory	7.17	32.39	2.92	5.79	46.03	2.40	3.42	4.73	0.00	13.96
Labor / Education / Health	2.87	11.63	83.75	5.58	6.35	6.41	10.96	16.62	23.40	40.74
Macroeconomics	1.79	14.12	5.36	3.37	4.13	4.41	64.38	6.89	6.38	11.11
Total	100	100	100	100	100	100	100	100	100	100
N	279	602	1,231	950	315	499	146	740	47	351

Note: The total number of observations is 5,160.

Table A3: Categorization of Countries into Former Territories

Category	**Countries**
British	Bangladesh, Cyprus, Egypt, Ghana, Hong Kong, India, Jamaica, Kenya, Malaysia, Nigeria, Pakistan, South Africa, Sri Lanka, Sudan, Uganda, Zambia, and England
French	Algeria, Benin, Cameroon, DR Congo, Côte d'Ivoire, Guinea, Haiti, Lebanon, Madagascar, Morocco, Niger, Senegal, Vietnam, and France
Spanish	Argentina, Bolivia, Chile, Colombia, Costa Rica, Dominican Rep., Ecuador, Guatemala, Honduras, Mexico, Paraguay, Peru, Philippines, Uruguay, Venezuela, and Spain
Portuguese	Angola, Brazil, Cape Verde, Guinea-Bissau, Mozambique, Timor-Leste, Macau, and Portugal

Table A4: Examples of Rationale for Classification of Countries with Multiple Colonial Histories

Country	**Dominant legacy and rationale**
Cameroon	Most of the territory was governed by France and retains French law and language in universities; only the smaller Western strip was under Britain, so it is coded as French.
Honduras	Although Britain controlled the Bay Islands at some point, Honduras' legal code, language of instruction, and university system derive from Spanish rule, so it is coded as Spanish.
Ghana	Although there was Portuguese rule historically, modern Ghana's courts, universities, and civil service were created during British rule (1874–1957); so it is coded as British.
India	Goa (Portuguese until 1961) covers less than 2 percent of India's population; legal and educational systems derive from British rule; so India is coded as British.
Morocco	France controlled the administration; universities follow the French model; so it is coded as French.

Table A5: Hausman Test for Violation of IIA

Omitted	**Chi-squared**	**df**	**P $>$ chi^2**	**Evidence**
Others (P, A, K, M, R, Y, Z)	0	1	1	for H_0
Econometrics (C)	0	1	1	for H_0
Micro (D)	0	1	1	for H_0
Labor/Health (I,J)	0	1	1	for H_0
Macro/Finance (E,G)	0	1	1	for H_0
IO (L)	0	1	1	for H_0
Environ & Agric (Q)	0	1	1	for H_0
Public (H)	0	1	1	for H_0
Dev/Growth/Int (O,F)	0	1	1	for H_0
Econ History (B,N)	-0.025	128	1	for H_0

Note: The table reports results from Hausman tests of the Independence of Irrelevant Alternatives (IIA) assumption in the multinomial logit model, using the field specialization variable harmonized from JEL codes and topic modeling output. The null hypothesis (H_0) is that IIA holds - that is, the relative odds of choosing between any two outcome categories are independent of the presence or absence of other alternatives. Failing to reject H_0 indicates no evidence of violation of the IIA assumption.

Table A6: Regression Results (based on dissertations with a single JEL field)

	Others	**Econometrics (C)**	**Micro (D)**	**Labor/Health (I,J)**	**Macro/Finance (E,G)**	**IO (L)**	**Environ & Agric (Q)**	**Public (H)**	**Dev/Growth/Int (O,F)**	**Econ Hist (B,N)**
Ref: North America										
Europe	-0.03***	0.01	0.00	-0.08***	0.12***	0.01	-0.05***	-0.01	0.04**	0.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)
LAC	-0.04**	0.09**	-0.18**	-0.01	0.20**	-0.06	0.01	0.00	0.01	-0.01
	(0.02)	(0.04)	(0.09)	(0.13)	(0.10)	(0.05)	(0.05)	(0.01)	(0.08)	(0.00)
South & Central Asia	-0.03*	0.02*	0.00	-0.08***	0.10***	-0.03*	-0.02	-0.02	0.07***	-0.01*
	(0.02)	(0.01)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
East Asia	-0.01	0.05***	0.01	-0.13***	0.11***	0.00	-0.04***	-0.02***	0.05***	-0.01***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)
Sub-Saharan Africa	-0.02	-0.01	-0.13***	-0.03	0.12***	0.00	0.02	-0.01	0.15***	-0.08***
	(0.02)	(0.03)	(0.05)	(0.04)	(0.03)	(0.02)	(0.03)	(0.01)	(0.03)	(0.02)
MENA	-0.05**	0.05***	0.01	-0.11***	0.11***	-0.01	-0.05*	-0.01	0.06**	-0.01*
	(0.02)	(0.01)	(0.03)	(0.04)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.00)
Year fixed effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

*Note: The table reports average marginal effects from multinomial logit regressions of field specialization on region of origin and country-level characteristics. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All specifications include controls for department rank, share of female faculty, and whether the university is public. Standard errors are clustered at the university level.*

Table A7: Regression Results (based on dominant field identified using topic modeling classification)

	Agriculture	**Development/ Growth/ Int.**	**Econometrics**	**Macro/Finance**	**IO / Game Theory (Micro)**	**Labor/ Health/ Education**
Ref: North America						
Europe	-0.04***	-0.02	0.05***	0.10***	0.00	-0.10***
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)
LAC	0.06	0.02	-0.03	0.03	-0.06	-0.03
	(0.06)	(0.08)	(0.07)	(0.11)	(0.08)	(0.13)
South & Central Asia	0.00	-0.04*	0.03	0.10***	-0.05*	-0.05
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
East Asia	-0.02*	-0.01	0.08***	0.09***	0.00	-0.14***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
Sub-Saharan Africa	0.06*	0.00	-0.01	0.05	-0.07	-0.04
	(0.03)	(0.02)	(0.04)	(0.04)	(0.05)	(0.04)
MENA	-0.02**	-0.01*	0.05**	0.07**	-0.02**	-0.06*
	(0.02)	(0.01)	(0.05)	(0.07)	(0.02)	(0.06)
Year fixed effects	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES

Note: The table reports average marginal effects from multinomial logit regressions of field specialization on region of origin and country-level characteristics. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *All specifications include controls for department rank, share of female faculty, and whether the university is public. Standard errors are clustered at the university level.*