Hertie School

The Evidence Interface

How policymakers encounter, engage with, and make sense of scientific knowledge

Berlin, July 2025

Sebastian Ramirez-Ruiz

Dissertation submitted to the Hertie School
in partial fulfilment of the requirements for the degree of

Doctor rerum politicarum (Dr. rer. pol.)

in the

Doctoral Program in Governance

Advisors

First advisor
Prof. Dr. Simon Munzert
Hertie School

Second advisor
Prof. Dr. Fabrizio Gilardi
University of Zurich

Third Advisor
Prof. Kevin Munger, PhD
European University Institute

zer, 2025; Klar et al., 2020). For instance, sharing findings on social media is associated with increased reach outside academic audiences (Côté and Darling, 2018) and impact metrics (Eysenbach, 2011; Peoples et al., 2016).

Crucially, interactions on social media produce digital traces that enable the large-scale analysis of public-facing behaviors. Prior research has shown that online networks often reflect offline social structures, and that digital behaviors can meaningfully reveal underlying 'analogue' preferences and dispositions (Dunbar et al., 2015; Barberá, 2015; He and Tsvetkova, 2023). In the context of political elites, such data have been used to study how legislators interact with constituents (Spierings, Jacobs, and Linders, 2019) and interest groups (Bunea, Ibenskas, and Weiler, 2025). The underlying logic in this work is that decisions to follow, reply to, or mention others on platforms like Twitter signal attention, interest, or perceived informational relevance.

Building on this logic, I argue that legislators' engagement with researchers on social media—such as following, mentioning, or replying—can be interpreted as public signals of attention to academic actors. These behaviors do not constitute direct evidence of evidence use or policy learning, but they offer observable indicators of legislators' willingness to acknowledge, amplify, or associate with scientific expertise in the public sphere. In this way, they reflect a broader orientation toward knowledge and the symbolic value of engaging with expert communities.

This setting offers a unique opportunity to compare patterns of elite–expert engagement across contexts. The accounts operate within a unique framework of platform-imposed behaviors and embeds all users within a shared public network, making the behavioral signals comparable. Moreover, different types of interactions on platforms like Twitter carry distinct social meanings—varying in terms of visibility, cognitive cost, and symbolic weight (Metaxas et al., 2015; Wojcieszak et al., 2022). Examining these behaviors can shed light on the strategic, partisan, and situational factors that shape when and how political elites publicly signal engagement with scientific expertise.

Data, methods, and empirical setting

The data for this study were collected between November and December 2022 and center on the Twitter activity of elected legislators in office during that year. The dataset spans 12 democracies with widespread Twitter adoption, including the three largest in both South and North America (Argentina, Brazil, Colombia, Mexico, Canada, and the United States), as well as six in Europe (France, Germany, Ireland, Italy, Spain, and the United Kingdom).

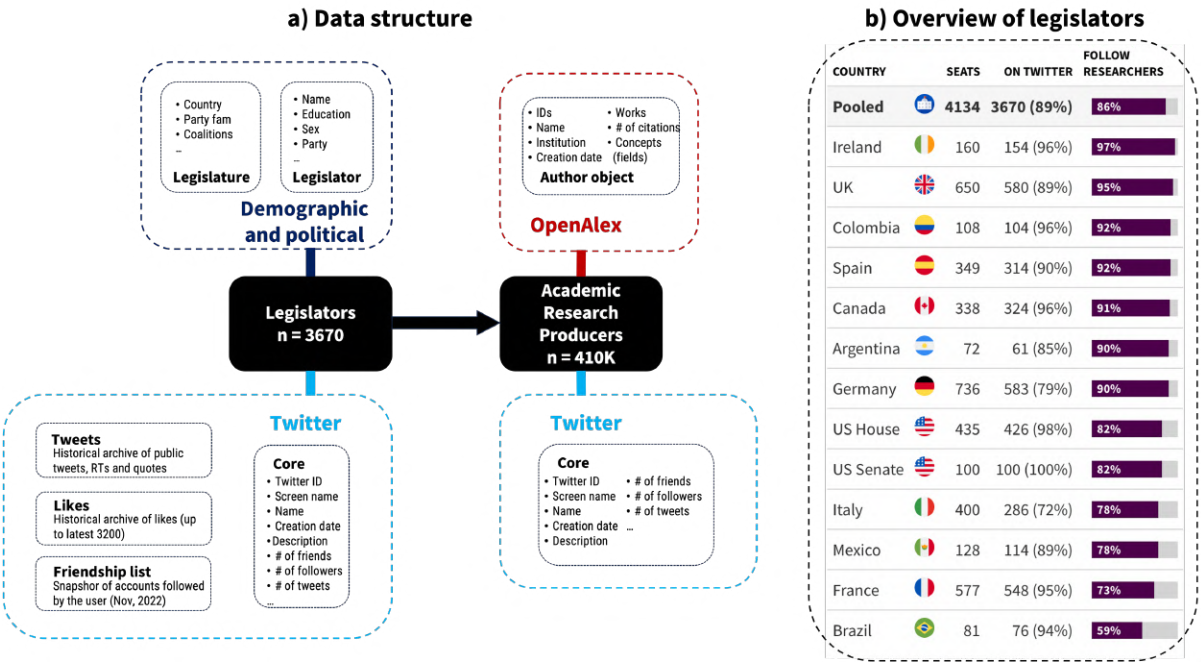


Figure 1: Study population and data structure. Panel a presents an overview of the data structure and connections between the legislator and academic researcher data sources. Panel b provides a summary of the legislators included in the study.

This cross-national scope provides a broad comparative foundation for analyzing patterns of interaction between legislators and academic researchers on social media.

To examine the relationship of legislators with academic researchers on social media and answer the three overarching questions, I match an original dataset containing the historical Twitter behavior of 3,670 legislators from 12 different countries with a newly compiled database of academic researchers on Twitter.

On the legislator-end, I collect all public posts (≈20M), a list of their followed accounts (≈2.6M), and the historical archive of liked posts (≈6.5M) by lawmakers from 12 democracies in Western Europe, North and South America actively in office during 2022. Additionally, I supplement the Twitter data with legislators' demographic and political information.

On the academic researcher-end, the starting point is a public database of 410K researcher Twitter and OpenAlex IDs (Mongeon, Bowman, and Costas, 2023). These accounts were identified by Mongeon et al. using a high-precision matching approach that connects Twitter profiles to researchers who shared links to scholarly work on Twitter, based on data from the Crossref Event Data release (January 2022). I collect information from their Twitter profiles, linking them to their respective researcher entries on the open index of scholarly work, OpenAlex (Priem, Piwowar, and Orr, 2022). I extract background information about the academics, such as their scientific field.

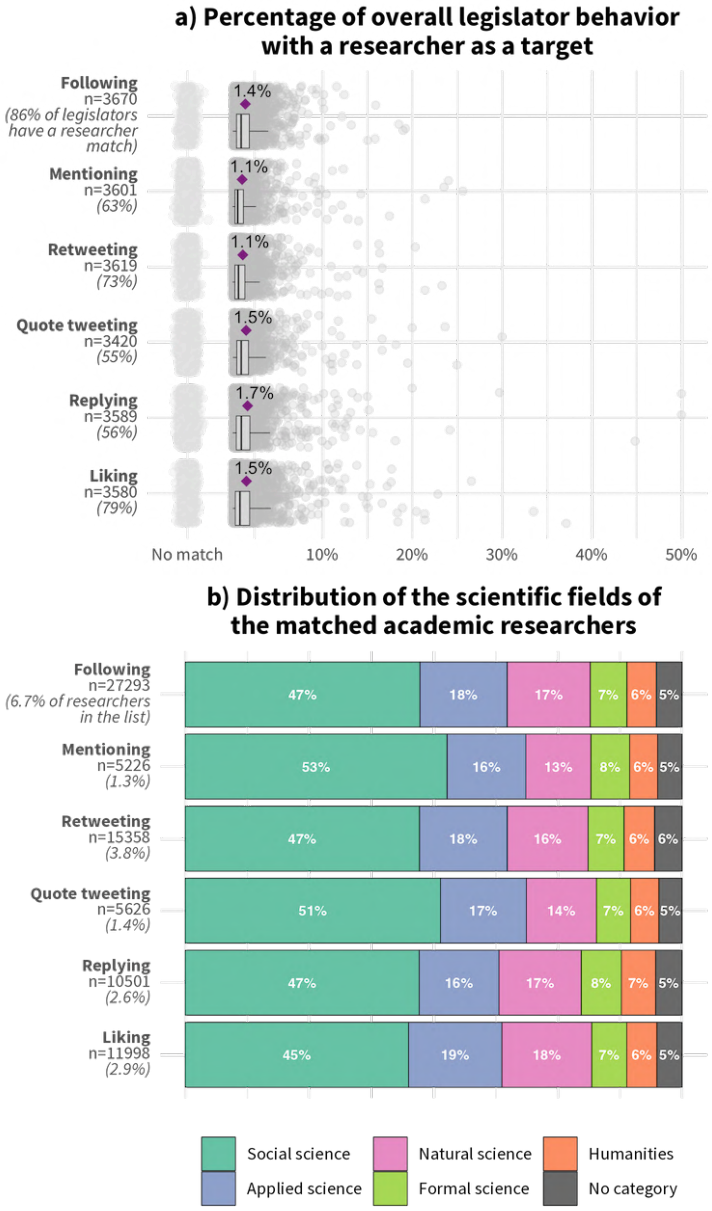


Figure 2: Overview of legislator Twitter behaviors targeted at academic researchers. In Panel A, dots represent individual legislators. The auxiliary information on the labels present the number of legislators who engage in the behavior and percentage who establish a researcher link at least once. Their position of the x-axis convey instances where events imply a legislator-researcher link as a proportion of the legislator totals. The purple rombi present the pooled legislator average. In Panel B, the auxiliary information denote the numbers and percentages of unique researchers from the original list identified as targets of legislators' Twitter events.

behaviors). However, following is concentrated amongst only a portion of accounts from the list (roughly 93% of the researchers are not in the legislators' Twitter 'radar').

The researcher list contains a number of power users with unusually strong social media followings, for example: @DrTedros (1.9M). In total, the dataset contains around 20 accounts with more than 1M followers and 500 with more than 100K. The finding remains stable when trimming the top 1% of followed accounts (keeping only researchers with less than 16K followers). Around 84% of the legislators follow at least one account from the restricted list

and the median legislator follows 6 users. Figure 2a presents an overview of the legislator-researcher Twitter events as a percentage of the total events for each behavior.

It is worth noting that the legislator accounts are highly active content-generators on Twitter. Approximately 60% of legislators have a daily posting rate, and a substantial 91% post on average once a week. In spite of the large volume of content production by legislators, only a small fraction can be linked to researchers. For instance, although about 73% of the legislators with retweets has ever retweeted a research producer, the median legislator has engaged in this behavior 4 times. Behaviors such as mentioning an original posts, replying to, and quote tweeting are even rarer with slightly more than half of the legislators in the sample engaging in them and the median legislator doing it only once.

The data availability of legislator 'likes' is constrained by Twitter API limits, which allowed to retrieve latest 3,200 posts liked by a user. This means that these numbers present a lower bound, since the full 'like' history is unobserved for roughly 30% of legislators who presumably use the feature more often. Overall, 79% of legislators liked at least one post by an academic researcher. The median legislator liked 6 posts by researchers.

Another set of interesting patterns comes from exploring the researcher-end features. In Figure 2b, I present an overview of the proportional make up of the researchers legislators interact with by their scientific discipline. These data present a picture in which social scientists consistently account for about half of those researchers linked to legislators across behaviors. This is substantially larger than researchers in other fields. For instance, around 19% of listed political scientists are followed by at least one legislator, compared to 3.4% of listed biologists. The overall pool of researchers in the study operate in diverse branches, e.g., natural (32%), social (27%), and applied (24%) sciences. Still, social scientists in the list are about 3.5x more likely to be followed against a comparable group in absolute size, natural scientists, and account for almost 47% of all followed researchers in the pooled network.

All-in-all, these descriptive analyses suggest that legislators do follow and engage with researchers in the 'digital wild'. Still there is an important qualification to these findings. These instances are highly uncommon, accounting for slightly more than 1% of legislator Twitter events of each behavior. *SI Appendix, Figure B2* provides some evidence of this qualification by analyzing the structure of the following lists of other groups—specifically, political and science journalists, as well as a random sample of researchers. The networks of these users can serve as a benchmark as their professional incentives could also make them likely to follow and engage with researchers. The comparison highlights that while

tors with research qualifications themselves were more likely to follow and engage with academic researchers. Additionally, academic researchers represent a higher proportion of these behaviors in the platform for legislators with research backgrounds compared to their non-researcher counterparts. The patterns attached to these legislators with higher educational qualifications can stem from their knowledge and interest about science, as well as the personal networks they created in the process of obtaining their degrees.

Furthermore, researcher producers in the list consistently represent a higher proportion of the networks and behaviors of younger, "digital native" and more popular legislator accounts. Notably, there were no observable differences between male- and female-identifying legislators.

Can legislators' inclination to follow and engage with researchers change? Evidence from a global pandemic. The early stages of the COVID-19 pandemic, marked by unfamiliarity and uncertainty, provide a scenario to explore the implications of exogenous shocks to the demand of specific types of expertise on legislators' online behaviors.

A number of studies about the early reactions in politics to the pandemic have investigated shifts in public opinion formation and electoral outcomes (Bol et al., 2021; Leininger and Schaub, 2023), as well as changes in the content of politicians online communications (Kim et al., 2022; Guntuku et al., 2021; Engel-Rebitzer et al., 2021). In this part of the study, I focus on Twitter as a marketplace of information and assess the temporal variation of legislators' Twitter behaviors towards academic researchers.

These analyses rely on the rise of the COVID-19 pandemic as an exogenous source of variation in the importance of scientific expertise. Given the magnitude, unexpectedness and salience of the pandemic, I leverage this variation to assess the average observed changes in legislator behaviors between the pre- and during-COVID periods. I use January 30, 2020 as the cut-point for the periods. This date marks the declaration of COVID as a public health emergency of international concern by the World Health Organization, as well as coincides with increased public awareness of the virus and its risks (see *SI Appendix, Figure B3* for an overview of COVID-related search term popularity around the period).

Based on the literature on policy learning and diffusion indicating that experts with specialized knowledge can gain salience and wield influence in situations dealing with new and technically complex policy issues (Haas, 1992; Dunlop, 2017), I expect that at the dawn of the COVID pandemic, researchers with expertise pertaining to the crisis will see increases in engagement by legislators. To test this expectation, I look at the potential for changes in

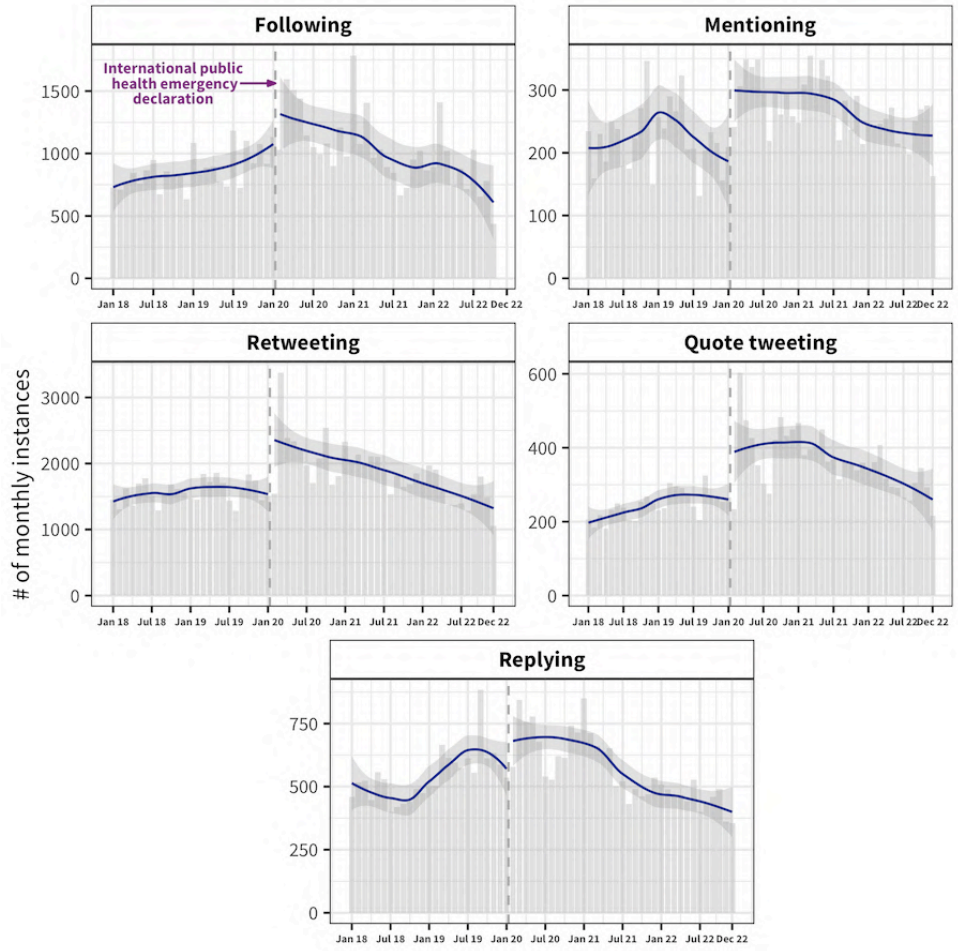


Figure 4: Distribution of the counts of monthly legislator-researcher instances by behavior between 2018 and 2022 (with LOESS smoothed moving average). The vertical dashed line marks the declaration of COVID-19 as an international public health emergency by the World Health Organization on January 30, 2023.

the rate of legislator-researcher links across Twitter behaviors between the pre- and during-COVID periods in the early stages of the pandemic.

The academic community swiftly reacted to the emergence of COVID-19. Surveys conducted among researchers in U.S. and European institutions reveal that approximately one-third of them shifted their focus to COVID-19-related research during the early phases of the pandemic across various disciplines (Gao et al., 2021). Even more, some research suggests that this scientific production quickly permeated policy circles, for example as source material for policy documents (Yin et al., 2021).

Figure 4 presents an overview of the distribution of the monthly legislator-researcher matches over the five-year period between January 2018 to December 2022. Across behaviors, the first half of 2020 is consistently amongst the periods with the highest number of instances in which legislators follow and engage with researchers during the time window. Most behaviors peak on March 2020, the month when COVID-19 was declared a pandemic.

I argue that the state of crisis, uncertainty, and urgency brought about by COVID-19 heightened demand for guidance from experts. Further, this demand is coupled with an increase in internet use and digital communication due to the spread of the virus. The context of the dawn of the pandemic make this a likely scenario for observing potential shifts in salience of different types of expertise translating into legislators' online behaviors.

That said, there can be some potential mechanisms at play that are not directly related to lawmakers' incentives, but reflect the broader impact that COVID could have had on science production and online behaviors more generally. For instance, potential changes in the supply of information or time spent online. It is possible that Twitter feeds contained more "scientific" content in the outset of COVID and people spent more time online. Still, it is not a logical consequence that politicians engage more when there is more science in their feed, rather one would expect that content needs to be relevant to translate into active behaviors.

Internet use increased during the pandemic (International Telecommunication Union, 2021). This seems to be also the case for legislators, who became more active on Twitter in the outset of the COVID period. There are observable increases across the range of platform-behaviors after January 2020, the only behavior that does not exhibit statistically significant changes is following (B = 3.4; 95% CI=[-3.7,10.7]; P=0.34). The volume of following new accounts is constant over the pre- and post-threshold periods (see *SI Appendix, Figure C12* for models exploring period differences across different bandwidths). That is to say, legislators' tweeting and engagement with tweets increased on average in the during-COVID period, yet the rate at which they created new links in their network remained unchanged.

Since the increase in online activity at the outset of the pandemic can obfuscate the analysis of absolute changes between the two periods, I focus my analysis on the estimated differences in the ratio of the probability of an event in the COVID period to the probability of an outcome in an pre-COVID period.

To achieve this, I estimate logistic mixed-effects models with country random effects to explore the likelihood that an observed legislator Twitter event targets a researcher in the COVID compared to the pre-COVID period. Figure 5a presents the marginal effects extracted from the models comparing a ±12-week time-window since the declaration of COVID-19 as public health emergency of international concern.

The results suggest that researchers were more likely to be at the receiving end of legislator behaviors on Twitter in the first months of COVID compared to the baseline period. The only behavior that showcases no discernible change is a behavior that requires scholars to

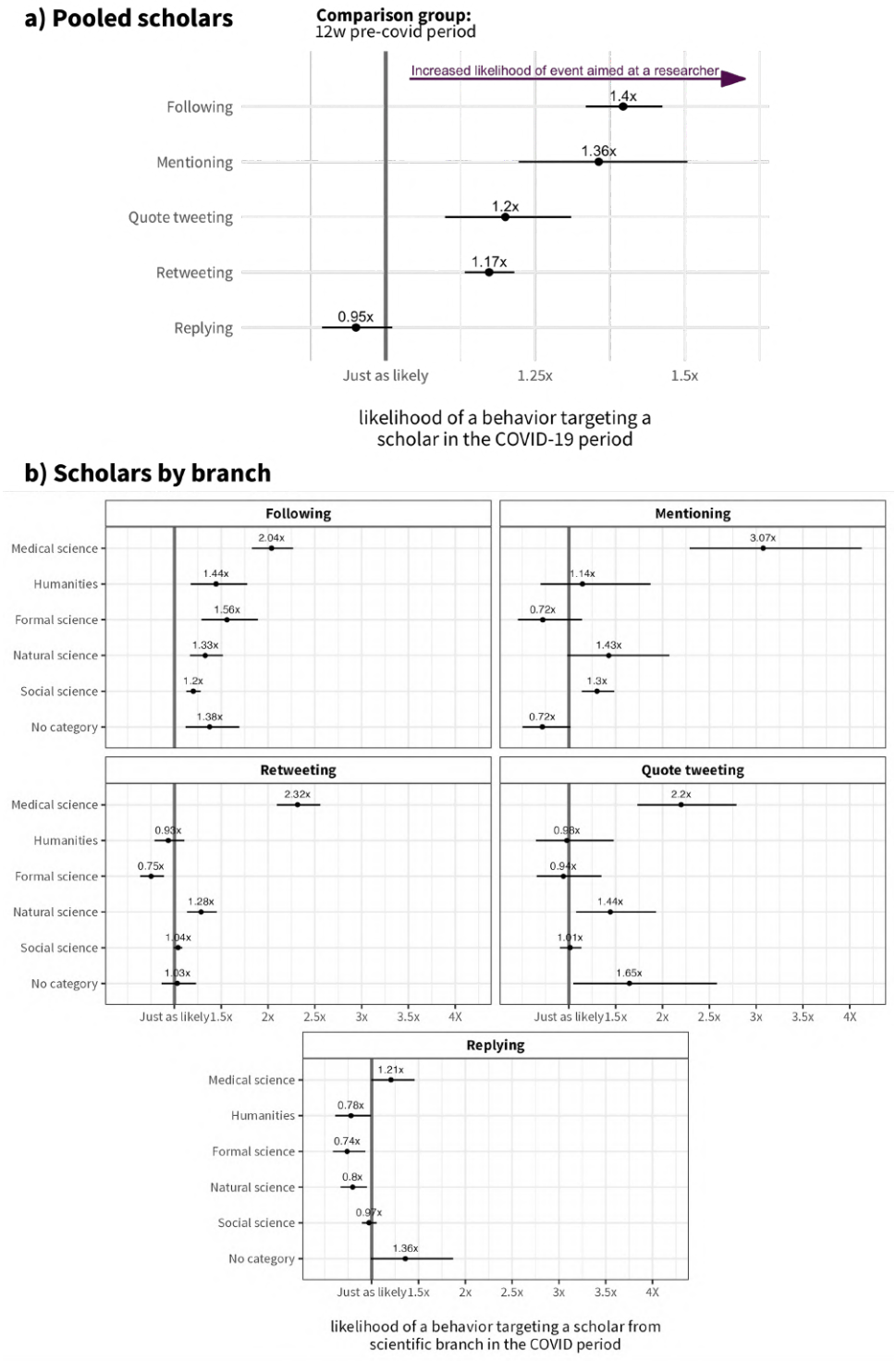
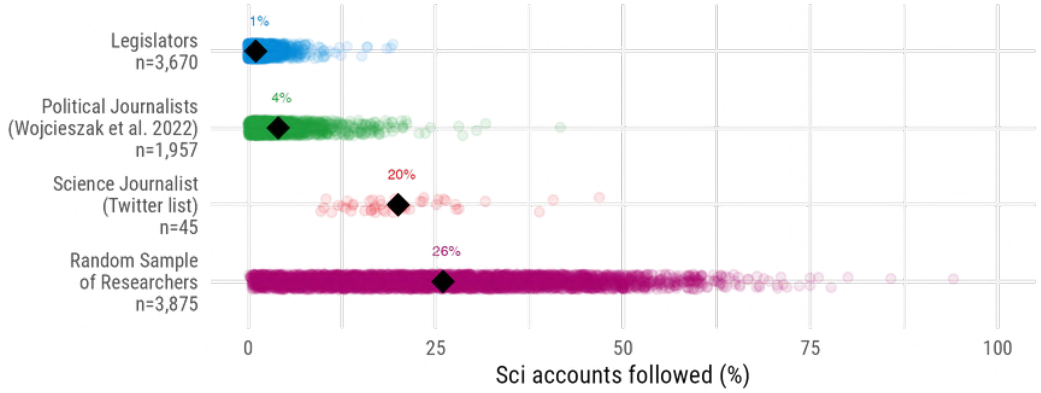


Figure 5: Marginal effects on following and engagement with academic researchers during the COVID versus pre-COVID periods with a ±12 week bandwidth. Results from a logistic mixed-effects models with country of legislature random effects. The estimates in the figure are relative risks representing the ratio of the probability of an event in the COVID period to the probability of an outcome in a pre-COVID period.

be first-movers by addressing legislators directly in their content in the first place, i.e., replying. For instance, new following edges in the 12 weeks after the declaration of COVID-19 as public health emergency of international concern period are 40% (B = 1.4; 95% CI=[1.33,1.46];

Figure B2: Percentage of overall Twitter followees matching users in the academic researcher list across groups.



Note. Each dot represents an individual user in each group. The black rombi represent the pooled group averages. The users for the *Political Journalist* group were extracted from (Wojcieszak et al., 2022). The *Science Journalist* users correspond to a curated science journalist list on Twitter (ID=1186721173222100994).

Figure B3: Google Topic interest of early COVID related terms

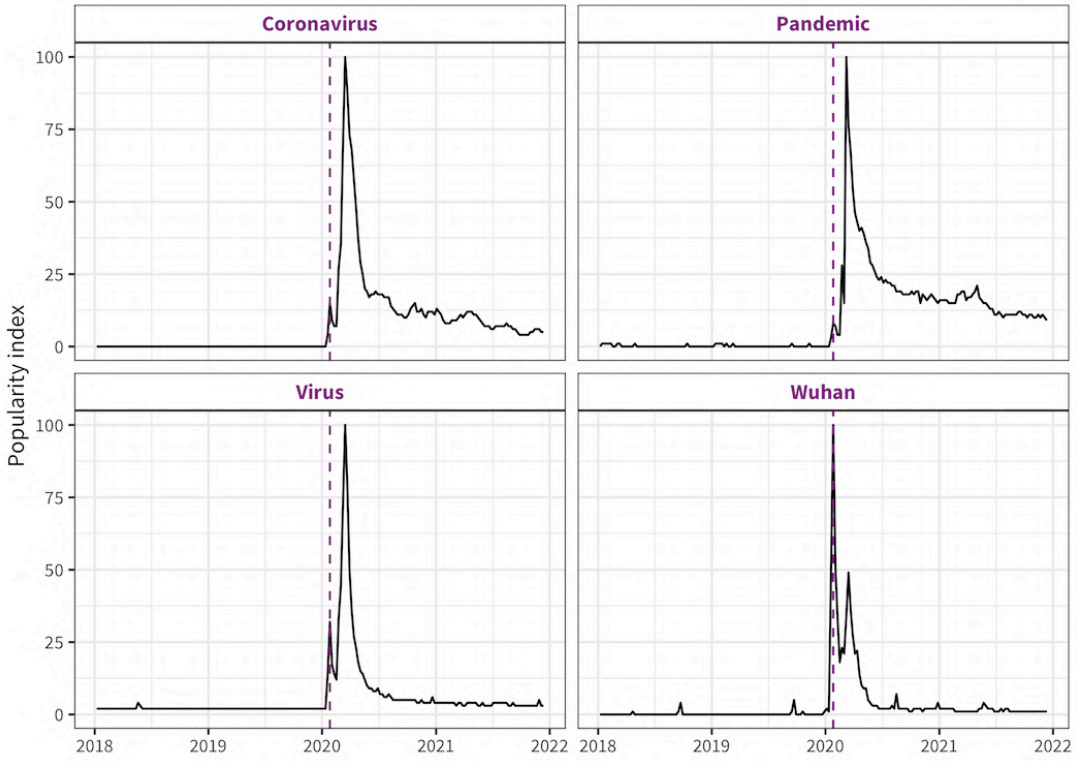


Table B4: Academic researchers with the largest increase in legislator followers following the declaration of COVID

Twitter handle	Name	Scientific branch	Last known affiliation	Base	Gained	Δ increase
@c_drosten	Christian Drosten	Natural science	Humboldt-Universität zu Berlin	52	102	1.96x
@DrTedros	Tedros Adhanom Ghebreyesus	Applied science	World Health Organization	24	52	2.17x
@uksciencechief	Patrick Vallance	Applied science	GlaxoSmithKline	35	50	1.43x
@devisridhar	Devi Sridhar	Applied science	University of Edinburgh	5	27	5.4x
@hendrikstreeck	Hendrik Streeck	No category	—	16	23	1.44x
@d_spiegel	David Spiegelhalter	Formal science	University of Cambridge	1	18	18x
@ronan_glynn	Robert J. Glynn	Applied science	Brigham and Women's Hospital	8	16	2x
@jasonleitch	J. Leitch	Social science	Scottish Government	5	15	3x
@GabrielScally	Gabriel Scally	Applied science	University of Bristol	1	15	15x
@globalhlthtwit	Anthony Costello	Applied science	UCL Institute of Child Health	3	15	5x
@JeremyFarrar	Jeremy Farrar	Applied science	Wellcome Trust	4	14	3.5x
@oriolmitja	Oriol Mitjà	Applied science	Fight AIDS Foundation	0	14	—
@anandMenon1	Anand Menon	Social science	Innovate UK	8	13	1.62x
@DrEricDing	Eric L. Ding	Applied science	Microclinic International	3	13	4.33x
@adamjkucharski	Adam J. Kucharski	Applied science	London School of Hygiene & Tropical Medicine	10	13	1.3x
@claire_ainsley	Claire Ainsley	Formal science	—	0	12	—
@mlipsitch	Marc Lipsitch	Applied science	Harvard University	4	12	3x
@CiesekSandra	Sandra Ciesek	Applied science	German Center for Infection Research	0	11	—
@CathCalderwood1	Catherine Calderwood	Applied science	Scottish Government	1	11	11x
@ASlavitt	Andrew Slavitt	Applied science	Centers for Medicare and Medicaid Services	4	10	2.5x
@pia_lamberty	Pia Lamberty	Social science	Johannes Gutenberg University of Mainz	0	10	—
@adam_tooze	Adam Tooze	Social science	Columbia University	4	10	2.5x
@GrimmVeronika	Veronika Grimm	Social science	University of Erlangen-Nuremberg	4	10	2.5x
@hans_kluge	Hans Kluge	Applied science	World Health Organization Regional Office for Europe	0	10	—
@miotei	Miguel Otero-Iglesias	Social science	Real Instituto Elcano	0	9	—

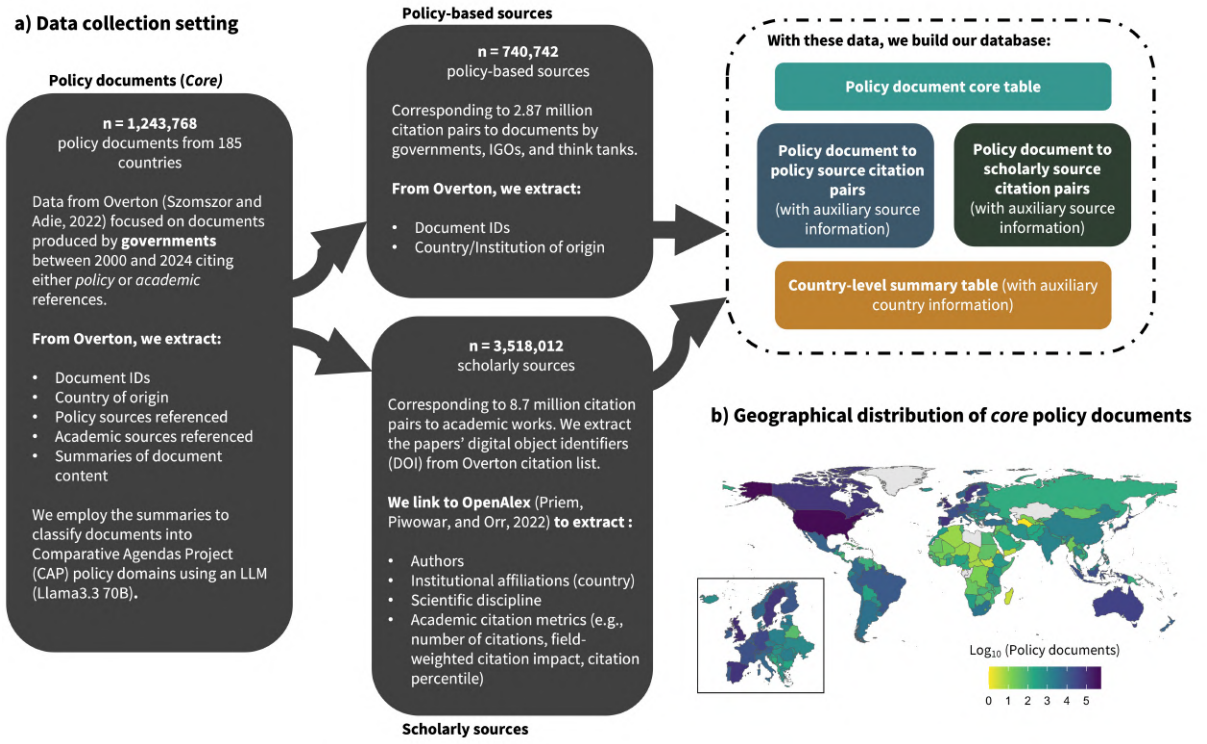


Figure 1: Study data collection and structure. Panel a illustrates the data sources, structure, and measures in our database. Panel b provides a summary of the geographical distribution of the *core* government policy documents.

edge needs. By doing so, we aim to provide a more comprehensive view of the expert knowledge landscape that informs government policymaking.

Results

Equipped with our novel data infrastructure, we conduct a systematic, large-scale analysis of citation dynamics within this global corpus of policy documents. Our empirical analyses are guided by three sequential questions: a) To what extent do governments rely on domestic versus foreign sources of evidence? b) Which countries' policy and scientific outputs are most frequently cited and therefore most visible? c) How do citation patterns vary across policy domains with differing knowledge demands? Together, these questions provide insight into whose knowledge is most prominently represented in global evidence-informed policymaking.

As introduced earlier, we distinguish between two broad categories of references: *policy-based* sources—documents from governments, intergovernmental organizations (IGOs), and think tanks—and *scholarly* sources, such as peer-reviewed journal articles and other scien-

tific outputs. This classification allows us to examine how different forms of knowledge are incorporated into government documents.

Policy-based citations in our dataset total approximately 2.8 million references to 740,742 unique documents. Most of these (80%) are authored by government agencies, followed by 11% from IGOs and 9% from think tanks or other organizations. On the scholarly side, government-authored documents cite academic publications over 8.7 million times, spanning roughly 3.5 million unique papers.

Looking across documents, we find that 52% cite only policy-based sources, 24% cite only scholarly sources, and the remaining 24% draw on both. This distribution suggests that while some documents blend diverse forms of expertise, the majority of documents in this corpus draw primarily from within the policy ecosystem.

General patterns of domestic and foreign citations in policy-based and scholarly references We first examine the types of evidence cited in policy documents to understand the reliance on domestic versus foreign sources of evidence. Because our data link citing and cited entities, we can trace the geographic flow of references—offering a view into how governments engage with knowledge produced within their own borders versus abroad.

For policy-based sources, we compare the origin of the cited document to the country authoring the citing document. For scholarly references, we identify institutional affiliations for 85% of cited works (3,012,209 in total) and classify papers as either: *Domestic* (all authors affiliated with institutions in the citing country), *mixed make-up* (at least one domestic affiliation), or *exclusively foreign* (no domestic affiliations).

A clear pattern emerges when it comes to policy-based sources. Governments in the Global North predominantly cite domestic sources, whereas those in the Global South more frequently reference foreign sources (see left panel of Figure 2). For example, only 13.2% of policy references in documents from African countries originate from national sources, compared to 73% in countries belonging to the U.N. Western European and Others Group (excluding the U.S.). This asymmetry underscores disparities in the transnational flow of policy-relevant information and the reliance on domestic expertise.

On the scholarly-end, the contrasts between Global North and South are less pronounced. That said, the proportion of science produced with at least one author associated with an institution within the country authoring the government document is larger in Western European and Others and Asia-Pacific groups compared to the other regions. In most geographies, the modal cited scholarly work is authored abroad—either solely by foreign-based

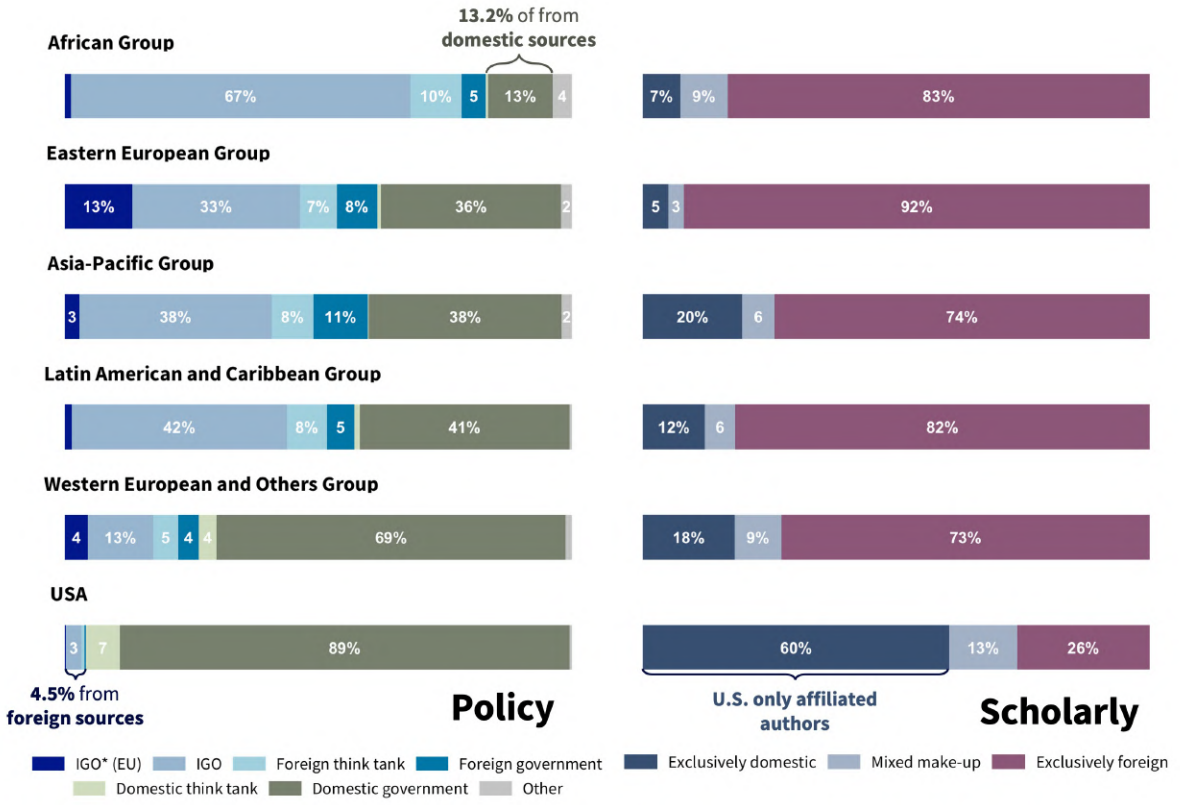


Figure 2: Composition of policy and scholarly source references by the policy documents across UN regional groups.

scholars or by teams spanning domestic and foreign institutions. This contrasts sharply with the United States, where 60% of cited papers are authored by researchers based at U.S. institutions.

While our analysis of domestic versus foreign references reveals notable differences in evidence use, it also prompts a broader question about visibility: which countries' outputs—both policy-based and scholarly—are most frequently cited beyond their borders? While a significant portion of foreign citations is directed toward documents from international organizations, many also cite policy materials produced by other national governments. In the academic realm, the majority of citations are dominated by researchers affiliated with institutions abroad. To explore these patterns of transnational influence, we now turn to an analysis of which countries' policy documents and scientific publications are most widely cited in global policymaking arenas, focusing on the references that "make it" across borders.

Global leaders in policy-based evidence reach and scientific visibility To quantify the international reach of government-produced evidence, we derive an *H*-index metric from the government-to-government citation matrix. This metric captures how often a coun-

try's individual policy documents are cited by others globally, providing insight into which nations have a larger footprint in the global policy document base.

Our analysis reveals significant regional differences in the citation of policy documents. Higher *H*-indices indicate countries whose policy documents are frequently referenced by other governments worldwide. The United States with 44 has the highest *H*-index, reflecting widespread reach and international recognition of its policy documents. The United Kingdom follows with an *H*-index of 32, while Germany, Australia, Canada, and several other European countries also exhibit strong citation patterns (see Figure 3a & c and Table A2).

On the other hand, a substantial proportion of countries—particularly those in the Asia-Pacific and African regions—are rarely cited in international policy documents. In fact, 30 countries (roughly 16% of the total) were never referenced in any policy documents from other nations during the study period. Most of these countries are classified as Least Developed Countries (LDCs) and Landlocked Developing Countries (LLDCs), highlighting unequal participation in global policy discourse.[1]

These patterns are robust across alternative metrics. For instance, they are further reflected in the centrality measures of the directed citation network between countries, as well as the inverse document frequency weighted country references and a country-centered *H*-index. Countries in the Global North consistently rank higher on betweenness, eigenvector, and PageRank centrality, illustrating their importance in global policy conversations (see *SI Appendix*, A.2)

As with policy citations, scholarly references are heavily skewed toward institutions located in the Global North. The United States, United Kingdom, Canada, Australia, and Germany are not only among the top sources of government-authored policy documents but also lead in cited academic research. Figure 3c illustrates this overlap. Notably, 17 of the top 20 countries for cross-border government citations also rank in the top 20 for scholarly citations by institutional affiliation. For instance, over 10,500 policy documents by other gov-

[1] In addition to these global metrics, we further examine regional variation in citation patterns using two complementary approaches. First, we run a series of logistic mixed-effects models to assess the probability that a reference targets sources from developed economies, neighboring countries, or within the same regional group. These models reveal differences across UN regional groupings, particularly in the extent to which countries in the Global South cite local or regional sources. Second, a correspondence analysis of intergovernmental citations illustrating how governments cluster based on shared citation behaviors. This exploratory technique highlights distinct citation logics across regions. For instance, the WEOG group of countries display relatively cohesive patterns, while other groups show more diffuse or externally oriented citation tendencies. Together, these analyses provide a more granular view of how geopolitical and regional dynamics might be at play in shaping patterns of knowledge exchange across policy systems (see *SI Appendix*, A.4).

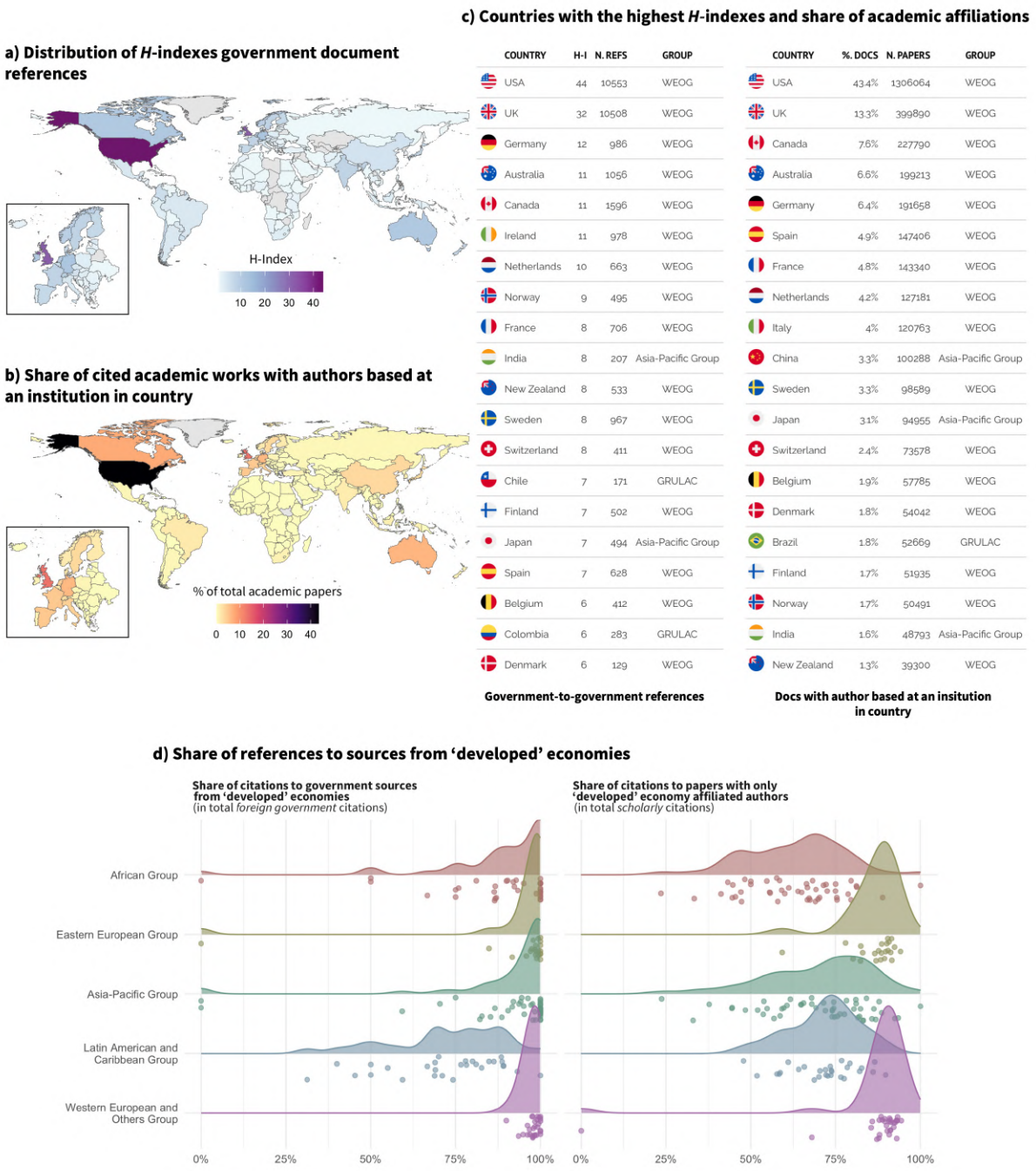


Figure 3: Overview of the distribution of reference metrics of government documents and academic works across countries. Panels a and b map the distribution of *H*-indexes and share of works with authors based at institutions in each country. Panel c presents the top-20 countries with the highest *H*-indexes and share of papers with authors linked to in-country institutions. Panel d illustrates the share of citations to foreign government and academic papers from 'developed' economies (dots represent individual countries in the regional group).

ernments in the corpus cite U.S. government publications, and 43% of all scholarly works referenced across the corpus include at least one author based at a U.S. institution.[2]

[2] An important qualification to these findings is that the academic research featured in policy documents generally aligns with the most prominently cited work within the scholarly community. We take this as a sign

Evidence use might vary across policy domains, but not its origins. Policy domains may differ widely in the types of expertise they require and the degree to which knowledge must be localized or specialized. Several scholars have proposed that distinct "cultures of evidence" may exist across different policy sectors (Lorenc et al., 2014). These cultures are thought to arise from factors such as the nature of decisions being made, policymakers' backgrounds and training, organizational norms, and the academic disciplines that inform each sector. However, when it comes to studies like ours—examining citations within policy documents—empirical research has so far focused almost exclusively on the climate change sector (Bornmann, Haunschild, and Marx, 2016). Our large-scale data enables a much broader and more detailed examination of evidence use across a diverse range of policy domains.

To explore these patterns, we classify each policy document using the Comparative Agendas Project (CAP) framework (Baumgartner, Breunig, and Grossman, 2019), which provides a standardized set of policy areas. We frame this as an automated text-annotation task, using a large language model (LLM) to assign each document's summary—provided by Overton—to the most likely CAP category. This method enables us to efficiently classify millions of documents at scale. The corpus contains varying numbers of documents across policy domains, and their distribution is relatively stable across world regions (see Table A6). Among these, documents in the Environment and Health domains are the most common, together accounting for roughly 38% of the corpus.

We find suggestive evidence that policy domains might indeed differ in their "cultures of evidence", particularly in the degree to which they rely on *policy-based* versus *scholarly* sources. To assess this, we categorize documents according to the type of evidence they cite: exclusively policy-based, exclusively scholarly, or a mix of both. Some domains—such as Housing and Government Operations—rely almost entirely on policy-based sources. Others—such as Macroeconomics, Agriculture, Public Lands, Health, and Environment—more frequently incorporate scholarly references, indicating a stronger orientation toward academic expertise (see *SI Appendix, Figure B2*).

We extend this analysis by examining the distribution of both foreign policy-based and scholarly references across domains. Figure 4 presents the results of four mixed-effects

that the academic research employed is also regarded highly within the academic environment. We analyze the field-weighted citation impact (FWCI)—a measure of how many citations a paper receives relative to the average for publications of the same year, type, and discipline—and the citation percentile (by year and subfield) of the referenced scholarly sources. Our analysis shows that the median scholarly work cited in policy documents receives 13 times more citations than comparable academic publications. These works also fall within the 92nd citation percentile of their field and year, suggesting that policy documents consistently draw on some of the most prominent scholarly research.

Policy documents across 185 countries predominantly rely on evidence from the Global North

Online Appendix

Contents

Appendix A Data and Statistical Analyses **81**
A.1 Data . . . 81
A.2 Government-to-government reference reach . . . 83
A.3 Classification of policy domains . . . 86
A.3.1 Validation Protocol . . . 87
A.4 Variation across regions . . . 90
A.4.1 Correspondence analysis . . . 92

Appendix B Supporting Figures and Tables **93**
B.1 Auxiliary information . . . 93

Appendix C Software statement **95**

Appendix A Data and Statistical Analyses

A.1 Data

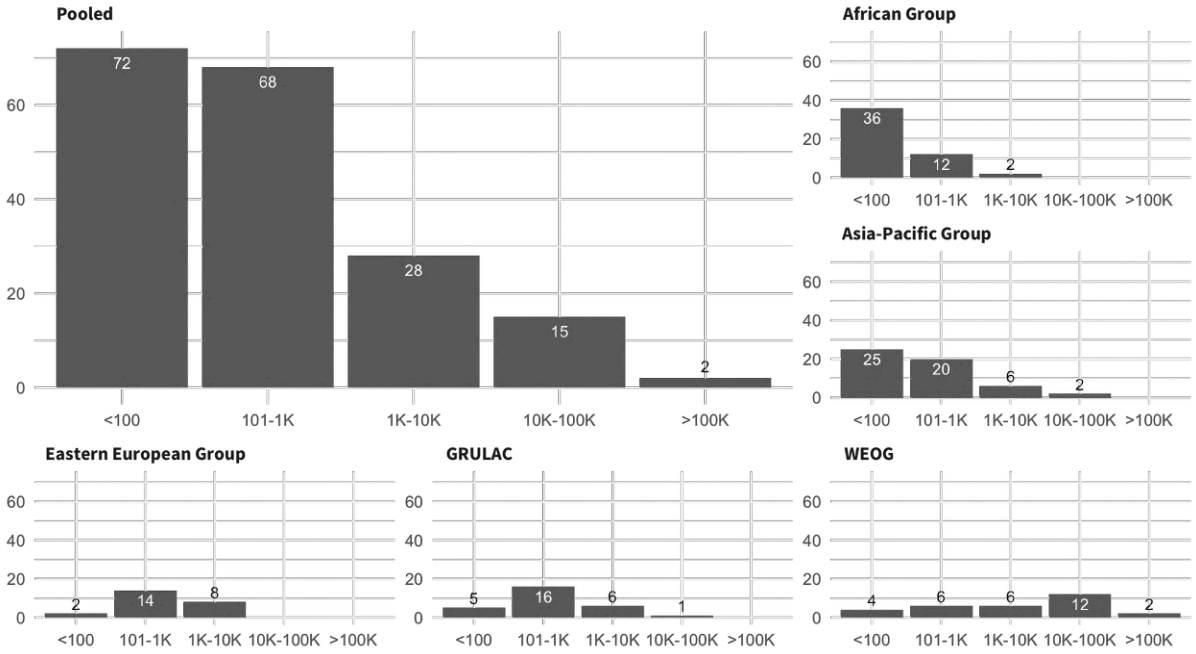


Figure A1: Overview of distribution of government authored policy documents. This figure shows the number of countries categorized by the total of policy documents collected, both in the full sample and grouped by UN Regional levels.

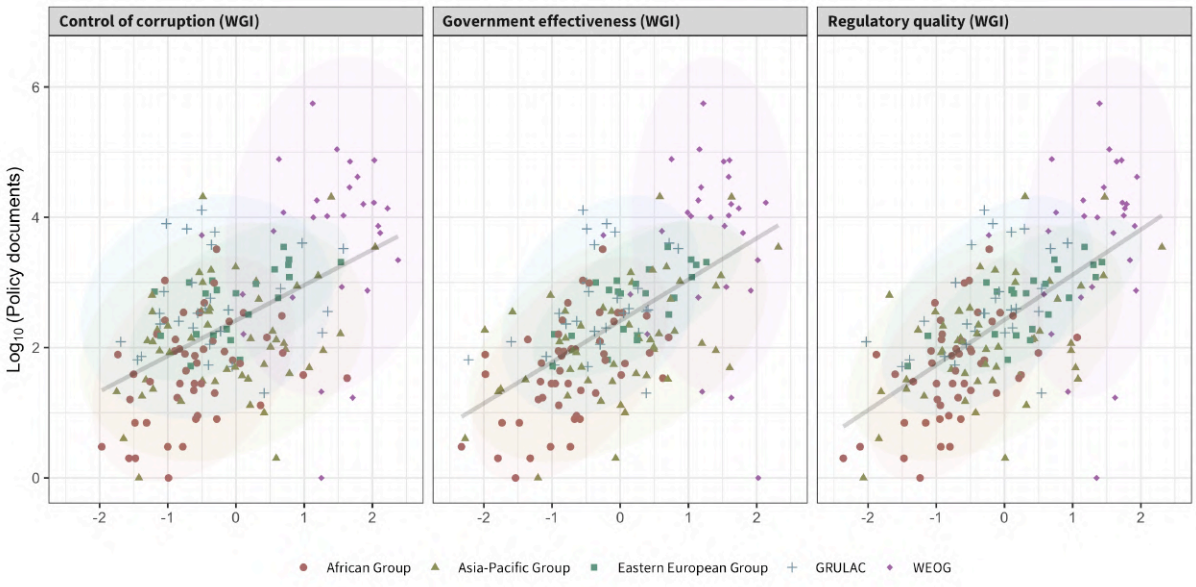


Figure A2: Relationship between Worldwide Governance Indicators (WGI) (*Worldwide Governance Indicators, 2024 Update, 2025*) and policy documents available across countries. Each dot represents individual countries, with ellipses at the UN Region-level.

A.4.1 Correspondence analysis

To further explore the structure of global evidence exchange, we conducted a correspondence analysis (CA) of intergovernmental citation patterns. We constructed a citation matrix that records how often a government in country A cited a document authored by a government in region B, producing a contingency table that summarizes citation flows among the 174 government citing countries and the 5 regions.

Correspondence analysis projects both citing and cited countries into a shared low-dimensional space, such that proximity reflects similarity in citation behavior. Countries that cite similar sources appear near each other in the space, while countries that are cited by similar sets of actors also cluster.

To visualize these relationships, we present a biplot, which jointly displays both the citing countries and the cited regions in the same coordinate space. In this plot, the spatial proximity between elements indicates similarity in citation patterns—that is, citing countries positioned close to each other tend to cite similar regions, and cited regions located near one another tend to be referenced by a similar set of countries. The biplot allows us to identify clusters of countries with aligned citation behaviors and to examine which regions attract broader versus more localized attention.

The results suggest the presence of regional citation clusters, with countries in Latin America and the Caribbean displaying particularly distinct citation patterns. Countries located near the center of the plot exhibit citation behaviors that are closer to the overall average, while those on the periphery demonstrate more differentiated or localized citation tendencies.

While exploratory in nature, this analysis offers additional evidence of variation in intergovernmental knowledge flows, and complements the main findings on global asymmetries in evidence use. The CA plots provide a visual summary of both citation practices and citation visibility, helping to contextualize broader patterns in the international exchange of policy evidence.

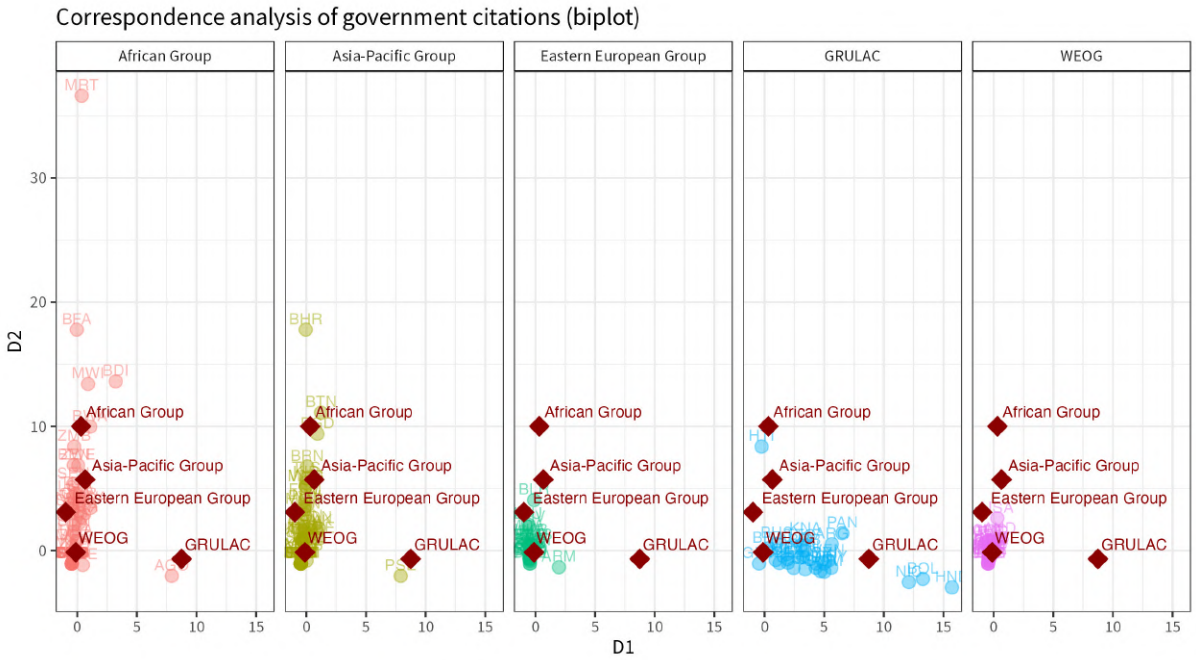


Figure A4: CA biplot.

Appendix B Supporting Figures and Tables

B.1 Auxiliary information

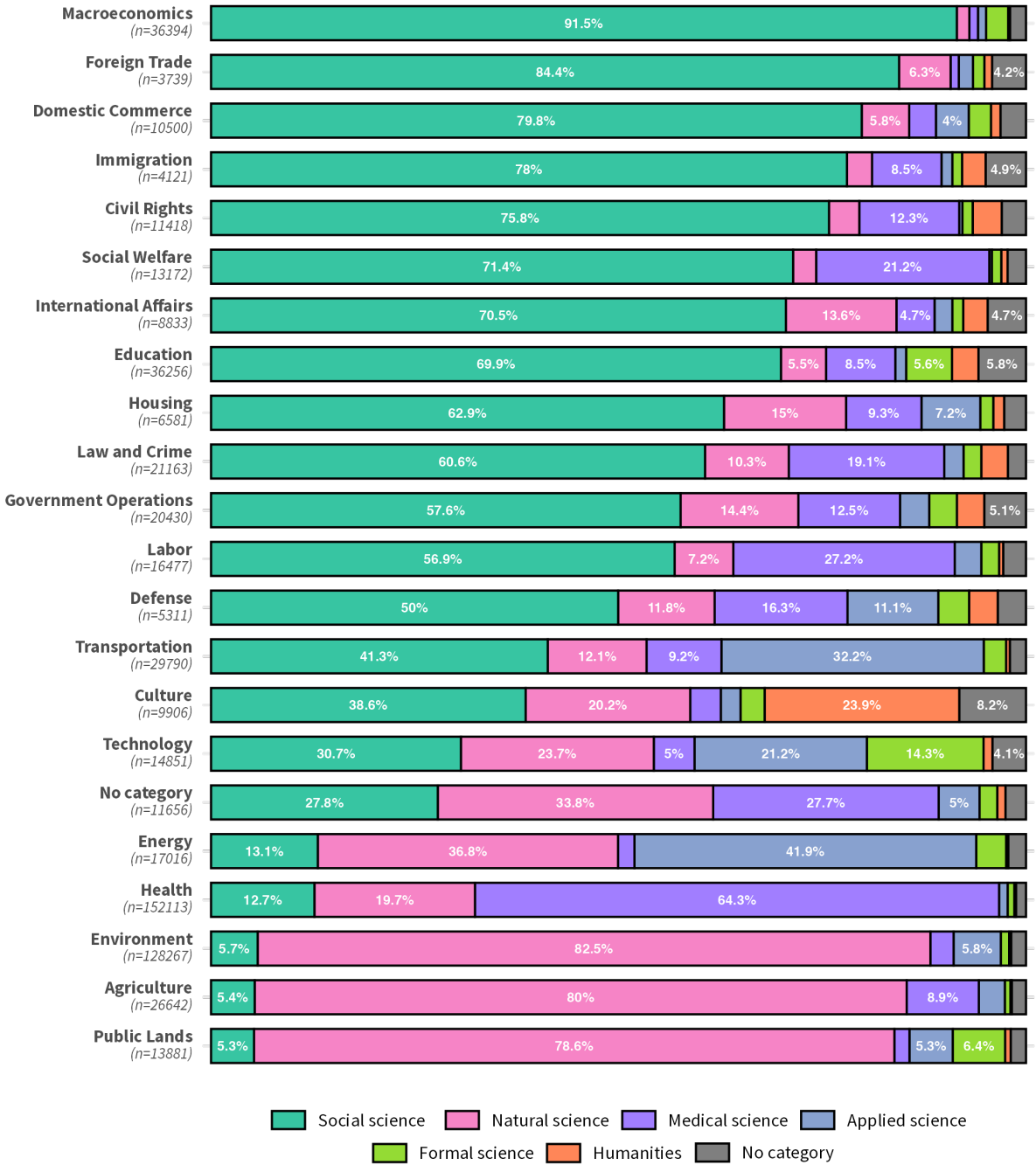


Figure B1: Share of references to scientific fields across policy domains.

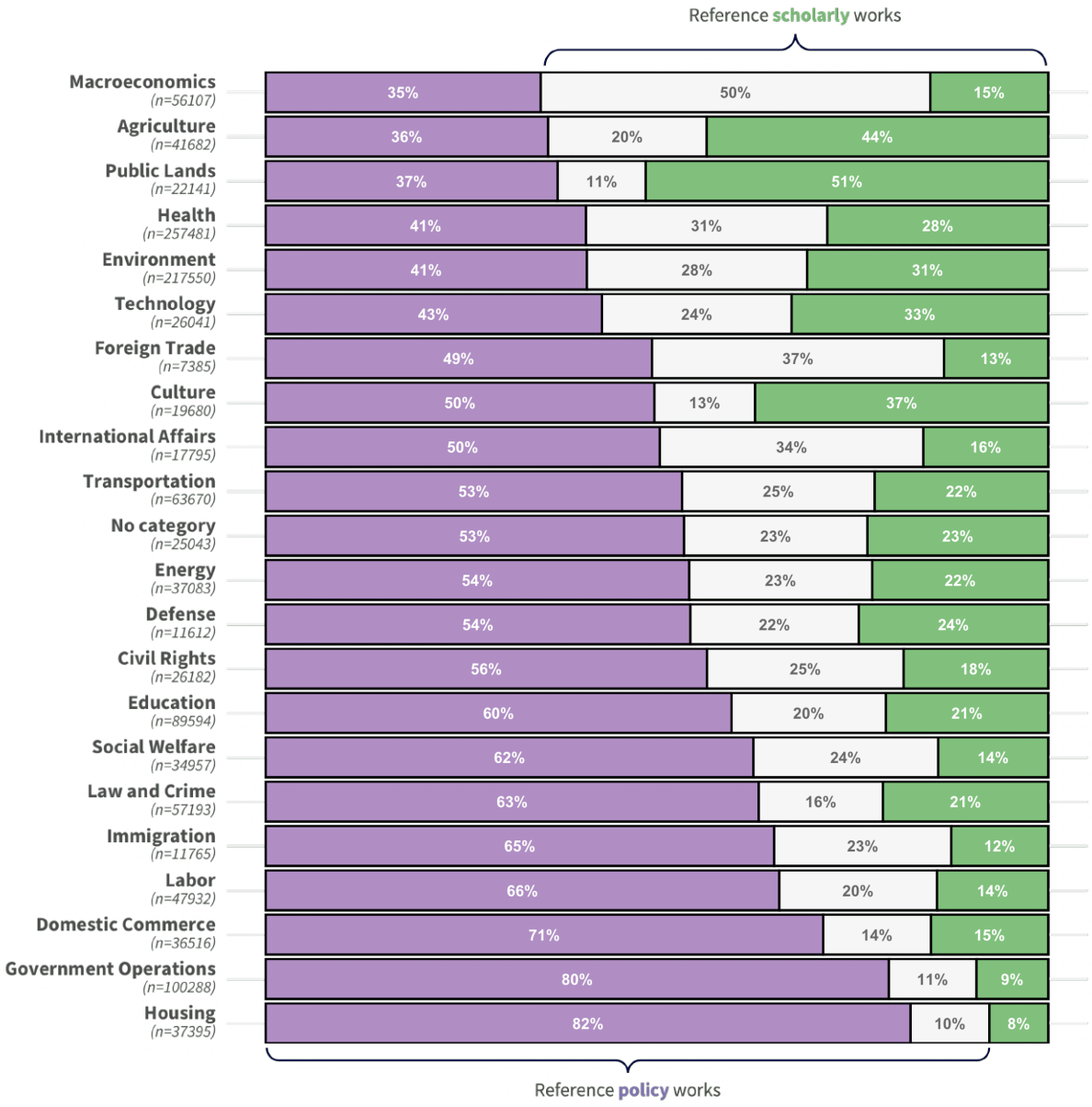


Figure B2: Share of policy documents citing policy, scholarly, or mixed sources across policy domains.

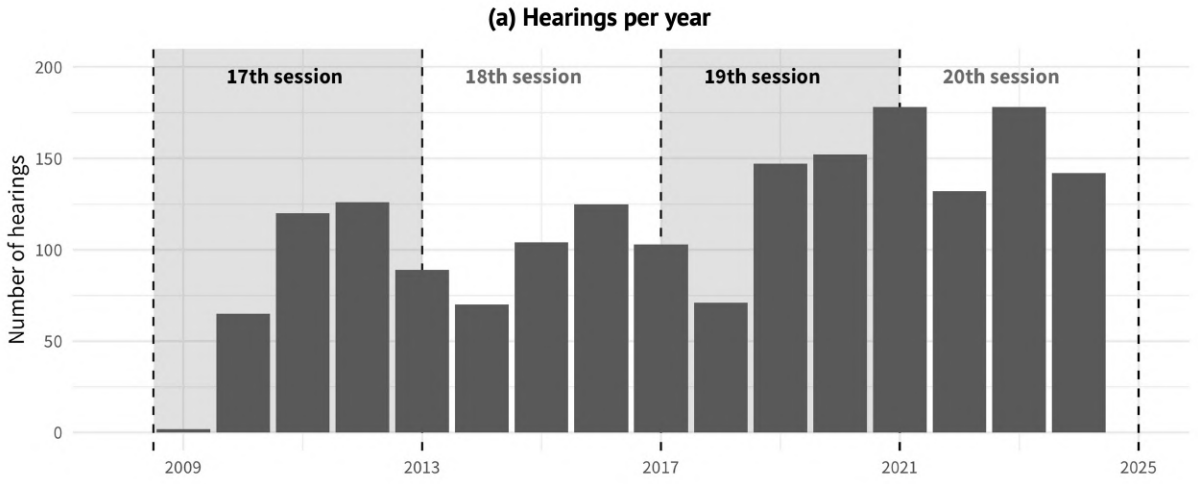
Table B1: Top 30 academic journals by policy document citations

Academic journal	Refs	% of docs
The Lancet	71437	0.9%
JAMA	61125	0.8%
New England Journal of Medicine	60913	0.8%
PLOS ONE	56756	0.7%
Environmental Science & Technology	52641	0.7%
BMJ	48513	0.6%
Science	46337	0.6%
American Economic Review	42465	0.5%
Environmental Health Perspectives	37291	0.5%
Nature	35007	0.4%
Science of The Total Environment	33564	0.4%
PEDIATRICS	32256	0.4%
Canadian Journal of Fisheries and Aquatic Sciences	31818	0.4%
The Quarterly Journal of Economics	29941	0.4%
Proceedings of the National Academy of Sciences	29754	0.4%
The Journal of Wildlife Management	29541	0.4%
Accident Analysis & Prevention	28108	0.4%
SSRN Electronic Journal	27798	0.3%
Journal of Political Economy	27745	0.3%
American Journal of Public Health	27413	0.3%
Journal of Monetary Economics	26785	0.3%
Econometrica	24865	0.3%
Open-File Report	23567	0.3%
Cochrane Database of Systematic Reviews	23058	0.3%
Marine Ecology Progress Series	22940	0.3%
Clinical Infectious Diseases	22819	0.3%
Biological Conservation	22479	0.3%
Chemosphere	21652	0.3%
Journal of Geophysical Research	21291	0.3%
The Journal of Finance	21032	0.3%

Appendix C Software statement

We used R version 4.4.2 (R Core Team, 2024) and the following R packages:

`ca` v. 0.71.1 (Nenadic and Greenacre, 2007)
`countrycode` v. 1.6.0 (Arel-Bundock, Enevoldsen, and Yetman, 2018)
`data.table` v. 1.16.4 (Barrett et al., 2024)
`dplyr` v. 1.1.4 (Wickham, François, et al., 2023)
`ggh4x` v. 0.3.0 (van den Brand, 2024)
`ggplot2` v. 3.5.1.9000 (Wickham, 2016)
`glmmTMB` v. 1.1.10 (Brooks et al., 2017)
`gt` v. 0.11.1 (Iannone et al., 2024)
`httr` v. 1.4.7 (Wickham, 2023)
`igraph` v. 2.1.4 (Csárdi et al., 2025)
`janitor` v. 2.2.0 (Firke, 2023)
`jsonlite` v. 1.9.1 (Ooms, 2014)
`marginaleffects` v. 0.24.0 (Arel-Bundock, Greifer, and Heiss, 2024)
`ollamar` v. 1.2.2 (Lin and Safi, 2024)
`openalexR` v. 1.4.0 (Massimo et al., 2024)
`purrr` v. 1.0.4 (Wickham and Henry, 2025)
`rnaturalearth` v. 1.0.1 (Massicotte and South, 2023)
`rnaturalearthdata` v. 1.0.0 (South, Michael, and Massicotte, 2024)
`tidyr` v. 1.3.1 (Wickham, Vaughan, and Girlich, 2024)
`xtable` v. 1.8.4 (Dahl et al., 2019)



(b) Overview of hearings with experts across committees

Committee	No. Hearings	Avg Year
Legal Affairs	258	17.2
Climate Action and Energy **(est. 2022)**	43	14.3
Health	209	13.9
Finance	191	12.7
Labor and Social Affairs	168	11.2
Internal Affairs and Community	167	11.1
Economic Affairs	106	7.1
Transport	100	6.7
Environment, Nature Conservation, Nuclear Safety and Consumer Protection	86	5.7
Housing, Urban Development, Building and Local Government **(est. 2018)**	30	4.3
Food and Agriculture	65	4.3
Family Affairs, Senior Citizens, Women and Youth	60	4
Human Rights and Humanitarian Aid	42	2.8
Budget	41	2.7
Tourism	40	2.7
Education, Research and Technology Assessment	38	2.5
Digital Affairs **(est. 2014)**	26	2.4
Sports	31	2.1
Cultural and Media Affairs	28	1.9
Economic Cooperation and Development	25	1.7
European Union Affairs	22	1.5
Defence	14	0.9
Foreign Affairs	9	0.6
Main Committee **(est. 2018)**	4	0.6
Petitions	1	0.1
Total	1804	120.3

Figure 1: Overview of public committee hearings with experts by year and committee between 2009 and 2024. The averages across committees are taken excluding 2009, since only 2 hearings were held immediately following the establishment of the 17th Bundestag.

standalone document, which is also annexed to the transcribed protocol of the hearings. I manually accessed and encoded the hearing materials containing the list of experts invited to testify in front of the committees. I extracted the names, affiliations, and roles of the experts. In total, the data set contains information for 11,437 expert-affiliation pairs[1].

[1]Expert-affiliation pairs refers to unique combinations of expert names and their listed institutional affiliations. The same individual may appear multiple times in the database if they participated in different hearings under different affiliations, or in different capacities (e.g., as a university researcher in one instance and as a representative of an advisory board in another).

Historically, the lists of experts contained the names of organizations and individual experts invited to the hearings. However, beginning with records from January 1, 2023, a procedural change introduced an additional layer of transparency: the Bundestag began publicly identifying which political party or committee member extended each expert invitation. This feature is incorporated into the database for hearings held after that date, allowing new inquiries into party-driven dynamics in expert selection.

To facilitate analysis, I classify experts according to their organizational affiliation. Categories include academic researchers, bureaucrats, interest groups and nonprofits, business representatives, government officials, and a residual "other" category. This classification was based on both the official institutional designations (e.g., eingetragener Verein—e.V. or Aktiengesellschaft—AG). While some edge cases required judgment calls—particularly for hybrid or multi-purpose organizations—the classification was applied systematically and documented. Researchers may refine these categories or supply their own classifications, which can be integrated into the database and contribute to its ongoing development through community-driven extensions.

Because the expert lists originate from official Bundestag documentation, the source material is highly structured. However, manual curation was essential to resolve inconsistencies in name formatting, institutional abbreviations, and metadata entries—especially for earlier sessions. The most common harmonization task involved standardizing institutional names across time. All records were flagged using a structured validation script and reviewed for consistency. Ambiguities were logged and manually resolved or annotated for

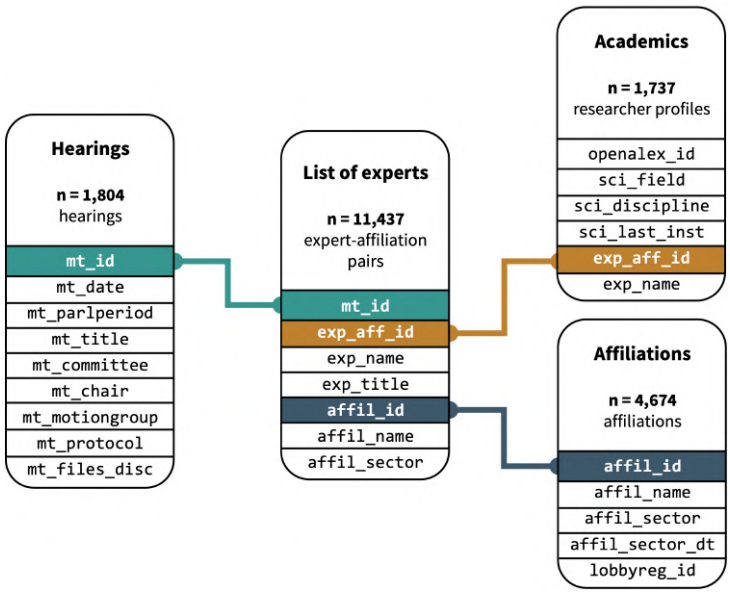
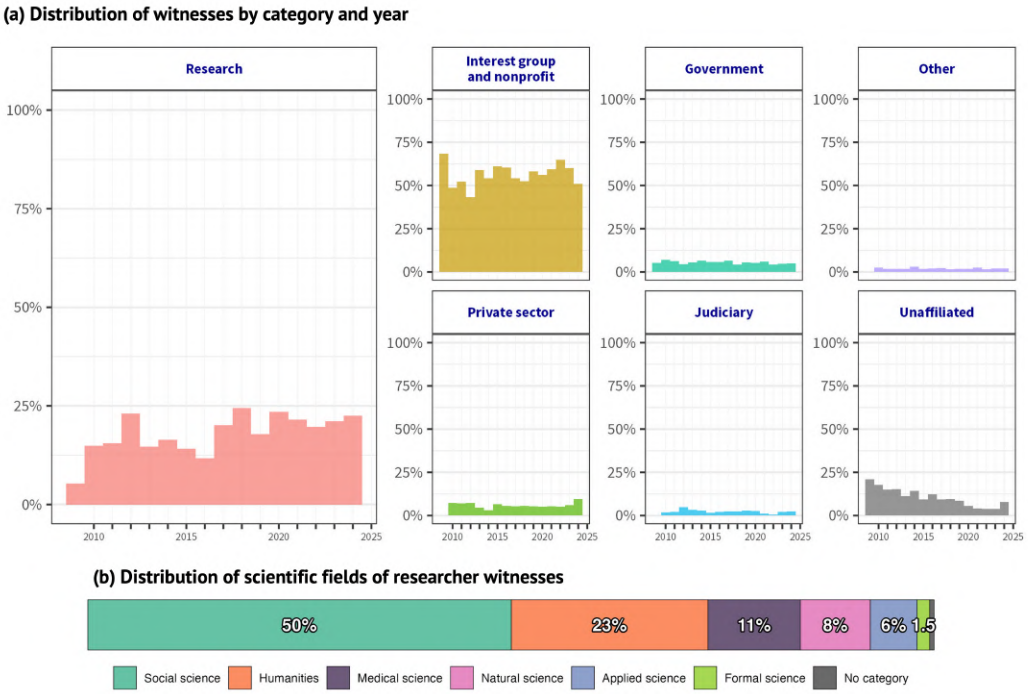


Figure 2: Illustration of the database structure.



(c) Overview of expert pool composition across committees

Committee	Experts	Res	Res (%)	Div indx.
European Union Affairs	146	100	68.5%	0.42
Foreign Affairs	51	32	62.7%	0.19
Petitions	7	4	57.1%	0.00
Budget	349	186	53.3%	0.44
Main Committee	40	18	45.0%	0.69
Defence	80	33	41.2%	0.65
Education, Research and Technology Assessment	285	102	35.8%	0.80
Internal Affairs and Community	1201	420	35.0%	0.61
Digital Affairs	205	68	33.2%	0.75
Legal Affairs	2112	595	28.2%	0.57
Human Rights and Humanitarian Aid	265	72	27.2%	0.30
Food and Agriculture	419	106	25.3%	0.63
Finance	1661	411	24.7%	0.44
Economic Cooperation and Development	142	35	24.6%	0.42
Health	1416	342	24.2%	0.68
Family Affairs, Senior Citizens, Women and Youth	536	129	24.1%	0.64
Economic Affairs	884	205	23.2%	0.57
Cultural and Media Affairs	209	42	20.1%	0.49
Env, Nature Cons, Nuclear Safety and Consumer Prot	704	128	18.2%	0.71
Sports	195	29	14.9%	0.44
Transport	741	104	14.0%	0.74
Climate Action and Energy	408	54	13.2%	0.63
Housing, Urban Development, Building and Local Government	268	35	13.1%	0.75
Tourism	258	31	12.0%	0.50
Labor and Social Affairs	1729	174	10.1%	0.66

(d) Distribution of researchers' by their lower level scientific domains

Primary topic	Field	N	Percent
Law	HUM	685	21.2%
Political Science and International Relations	SOC	376	11.7%
Sociology and Political Science	SOC	367	11.4%
Economics and Econometrics	SOC	237	7.3%
Strategy and Management	SOC	178	5.5%
General Health Professions	MED	139	4.3%
Accounting	SOC	131	4.1%
Finance	SOC	131	4.1%
Public Health, Environmental and Occupational Health	MED	75	2.3%
General Economics, Econometrics and Finance	SOC	41	1.3%
Electrical and Electronic Engineering	APP	35	1.1%
Renewable Energy, Sustainability and the Environment	APP	33	1.0%
General Agricultural and Biological Sciences	NAT	28	0.9%
Literature and Literary Theory	HUM	28	0.9%
Computational Mechanics	APP	26	0.8%
Epidemiology	MED	26	0.8%
Clinical Psychology	SOC	25	0.8%
Atmospheric Science	NAT	21	0.7%
Molecular Biology	NAT	21	0.7%
Plant Science	NAT	19	0.6%
Religious studies	HUM	18	0.6%
Immunology	NAT	17	0.5%
No category	NA	17	0.5%
Management Science and Operations Research	SOC	16	0.5%
Other		536	16.6%

Figure 3: Overview of research expert participation in Bundestag committee hearings.

all witnesses since 1960, while academic experts represent around 8% (Ban, Park, and You, 2023).

Panel C illustrates both the overall representation of academic researchers relative to other types of expert witnesses and the disciplinary diversity of the researchers appearing before different committees. The prominence of academic expertise varies substantially across policy domains. For example, researchers account for only around 10% of expert witnesses in the Committee on Labour and Social Affairs, whereas they constitute approxi-

mately 70% in the Committee on European Union Affairs. Other high-convening committees with relatively strong representation of researchers include Internal Affairs and Community, Legal Affairs, Finance, Health, and Economic Affairs, suggesting that certain domains consistently draw more heavily on academic input than others.

To capture the degree of disciplinary diversity within each committee, I compute the Gini–Simpson index, a common measure in ecological and sociological studies of diversity. This index reflects the probability that two randomly selected researchers invited by a given committee belong to different scientific fields. Higher values indicate greater disciplinary heterogeneity. The results reveal clear variation in how committees engage with scientific expertise. Committees such as Petitions, Foreign Affairs, and Human Rights and Humanitarian Aid exhibit highly concentrated disciplinary profiles, suggesting reliance on a narrow set of academic fields. In contrast, the Committee on Education, Research and Technology Assessment stands out for its high disciplinary diversity, drawing a broad spectrum of academics from different scientific branches. This pattern aligns with the committee's central role in overseeing research policy.

Beyond revealing overall patterns in the composition of expert witnesses, the BEWIT enables more fine-grained analysis of academic participants. By linking academic experts to their OpenAlex profiles, the database allows researchers to examine disciplinary trends and identify the types of scholarship that are more frequently represented in parliamentary hearings. For example, mirroring patterns observed in studies of social media engagement between scholars and legislators (Ramirez-Ruiz, 2025), social scientists are the modal group among academic witnesses, representing nearly half of all unique academic contributors. At the other end of the spectrum, scholars from the formal sciences (e.g., mathematics, computer science) are less represented, making up just 1.5%. Panel D offers a more detailed view, showing that legal scholars, political scientists, sociologists, economists, and management researchers constitute the top five contributing domains.

These patterns carry important implications for our understanding of how legislatures incorporate academic expertise. The consistent presence of researchers—especially in committees dealing with complex regulatory and legal issues—highlights the Bundestag's reliance on scholarly input in these sessions. At the same time, the differences in representation across disciplines and committees suggest that access to legislative fora is not evenly distributed across the academic landscape. Social scientists and legal scholars dominate the expert pool, while technical and formal sciences are less prominent. This distribution likely reflects the nature of the policy questions addressed in parliamentary committees,

many of which center on legal frameworks, social systems, and regulation. It may also suggest that some disciplines are more accustomed to engaging with legislative processes or more frequently consulted on issues with direct social implications.

Application 2: Does the composition of expert pools change when constituents can learn who invites whom to hearings?

The second application illustrates how the BEWIT can be used to examine the political dynamics shaping expert selection. Specifically, I assess whether a recent procedural reform introducing greater transparency in committee hearings is associated with shifts in the composition of invited experts. As of January 1, 2023, the Bundestag began publishing which parliamentary group invited each witness to testify. This reform offers a unique opportunity to evaluate whether making party sponsorship visible to the public affects the makeup of expert pools.

Supplementary table B3 details the distribution of invitations across parties. Roughly half of all expert invitations come from the CDU/CSU and SPD groups. Notably, the far-right AfD accounts for only 3% of invitations (despite holding 8% of seats at the time), and the newly formed BSW group, active only in the final months of the 20th Bundestag term, invited just five witnesses. Figure 4 provides a breakdown of expert types invited by each party. Across most parties, interest group representatives dominate the expert pool, followed by academic researchers. However, the AfD stands out, as nearly half of the experts it invited are unaffiliated, meaning they did not report any institutional association and appear as self-standing actors.

Understanding whose voices are amplified in parliamentary committees is essential to studying legislative representation and democratic legitimacy (Green, 2016; Geddes, 2018). Expert invitations might reflect ideological preferences, strategic priorities, and responses to outside actors. For example, recent research has highlighted persistent gender disparities among expert witnesses (Coil et al., 2024), while others show that interest groups are consistently overrepresented in the Bundestag's hearings (Dhungel and Linhart, 2014). Crucially, there is growing evidence that such imbalances are not just descriptive, but consequential. When constituents perceive that special interest groups exert disproportionate influence, it may erode trust in the legitimacy and fairness of legislative decision-making (Rasmussen and Reher, 2023, 2025).

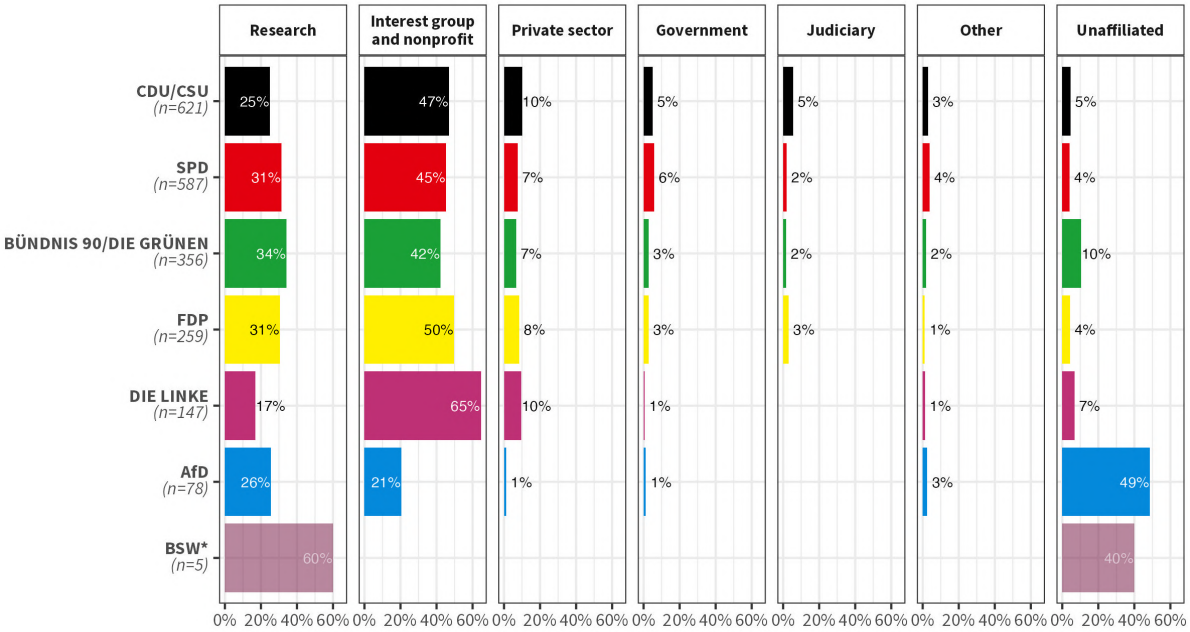


Figure 4: Expert composition by party group. This figure shows the distribution of expert types invited by each parliamentary group in hearings where the origin of invitations was publicly disclosed (i.e., following the change in protocol practices in January 2023). *The BSW group was officially formed in February 2024.

To evaluate whether the transparency reform correlates with changes in the composition of experts, I estimate a logistic regression with committee fixed effects. Specifically, I model the presence of members representing special interests. Figure 5 presents the average marginal effects for the full model and separately for the committees that held public committee hearings with experts during the 20th Bundestag. There are no estimates reported for the Defense and Petitions committees, as they did not convene such hearings during this period.

This analysis suggests that the introduction of the disclosure rule is associated with a 6-percentage-point decrease in the probability that an expert comes from an interest group or nonprofit. While this pre-post design cannot fully account for all confounding factors, the findings provide suggestive evidence that transparency around party sponsorship might modestly shift the types of experts who are selected to testify.

Together, these two applications underscore the analytical potential of the BEWIT for studying both the composition and politics of expert participation in legislative committees. Future research can build on these illustrations to investigate a wide range of questions surrounding information provision, epistemic authority, and the institutional design of legislative practice.

Domain and sub-domain	Item code	Item label	2PL		Percent correct	Median time (s)	N
			Disc (a)	Diff (b)			
Scientific Literacy							
Basic science	SEL04	Research credibility	1.73	-2.99	98.0%	14	175
Basic science	SEL03	Quality of evidence	1.91	-2.37	95.0%	18	170
Source reliability	SEL07	Trustworthy data	1.54	-2.39	94.0%	13	190
Source reliability	SEL08	Trustworthy sources	1.42	-2.49	94.0%	24	172
Scientific practice	SEL06	Scientific practice 2	1.65	-1.92	90.0%	21	176
Basic science	SEL01	Scientific hypothesis	1.61	-1.74	88.0%	10	189
Scientific practice	SEL05	Scientific practice 1	0.89	-2.44	85.0%	22	175
Basic science	SEL02	Scientific consensus	1.06	-0.03	50.0%	9	176
Statistical Literacy							
Machine learning models	SML05	ML and perpetuated bias	1.96	-2.68	97.0%	8	220
Machine learning models	SML07	ML application	1.04	-2.52	91.0%	10	210
Base rate	SML03	Base rate 1	1.40	-1.59	88.0%	39	160
Machine learning models	SML04	ML accuracy	1.27	-1.69	84.0%	9	234
Center and spread	SML01	Center and spread	0.82	-1.66	74.0%	36	163
Center and spread	SML02	Sample sizes and uncertainty	0.95	-0.66	65.0%	36	162
Machine learning models	SML06	ML and flawed predictions	0.82	-0.45	60.0%	29	193
Data Literacy							
Visual information	DVL06	Visual information 2	1.32	-2.97	95.0%	24	261
Numeracy	DVL01	Complementary probability 1	1.47	-1.97	90.0%	11	174
Visual information	DVL05	Visual information 1	0.86	-1.87	81.0%	48	241
Visual representation	DVL04	Visual data representation 1	0.99	-0.25	57.0%	33	235
Numeracy	DVL02	Complementary probability 2	1.65	-0.01	50.0%	31	182
Numeracy	DVL03	Percentages	0.93	0.23	48.0%	42	181
Problematic visuals	DVL07	Flawed visuals 1	1.14	0.50	41.0%	58	253
Problematic visuals	DVL08	Flawed visuals 2	0.85	0.89	37.0%	54	245
Causal Reasoning							
Experiments	CR04	Randomized control trials	0.80	-4.38	96.0%	10	160
Causal reasoning	CR05	Confounder	1.09	-1.90	85.0%	24	196
Valid causal conclusions	CR07	Study conclusions 1	2.09	-1.12	82.0%	26	181
Correlation	CR03	Learning from correlation	0.86	-0.97	69.0%	27	192
Causal policy	CR06	Causal effect of a policy	1.08	-0.65	64.0%	27	193
Correlation	CR02	Correlation	0.82	-0.49	60.0%	20	197
Correlation	CR01	Correlation vs. Causation	1.04	0.22	46.0%	10	168

Table 2: Overview of items in the INSPIRE inventory. This table presents item-level estimates from a two-parameter logistic (2PL) IRT model fitted to responses from N = 470 Prolific participants. For each item, the discrimination parameter (a) and difficulty parameter (b) are reported, alongside the sample share of correct answers and the median response time in seconds (measured as the time from item display to answer submission). A full list of items (including those that were excluded) and additional performance diagnostics are available in Online Appendix A.

We provide further diagnostic detail—including item coverage, response distributions, and model-based fit statistics—in Supplementary Appendix B.

Inventory performance and coverage

With the 30-item INSPIRE inventory now established, we proceeded to evaluate its overall psychometric performance and the range of ability it captures. Using the estimated parameters from a two-parameter logistic (2PL) IRT model on the item bank, we generated characteristic and information curves to visualize the inventory's properties across the latent ability (θ) score range. These are presented in Figure 1.

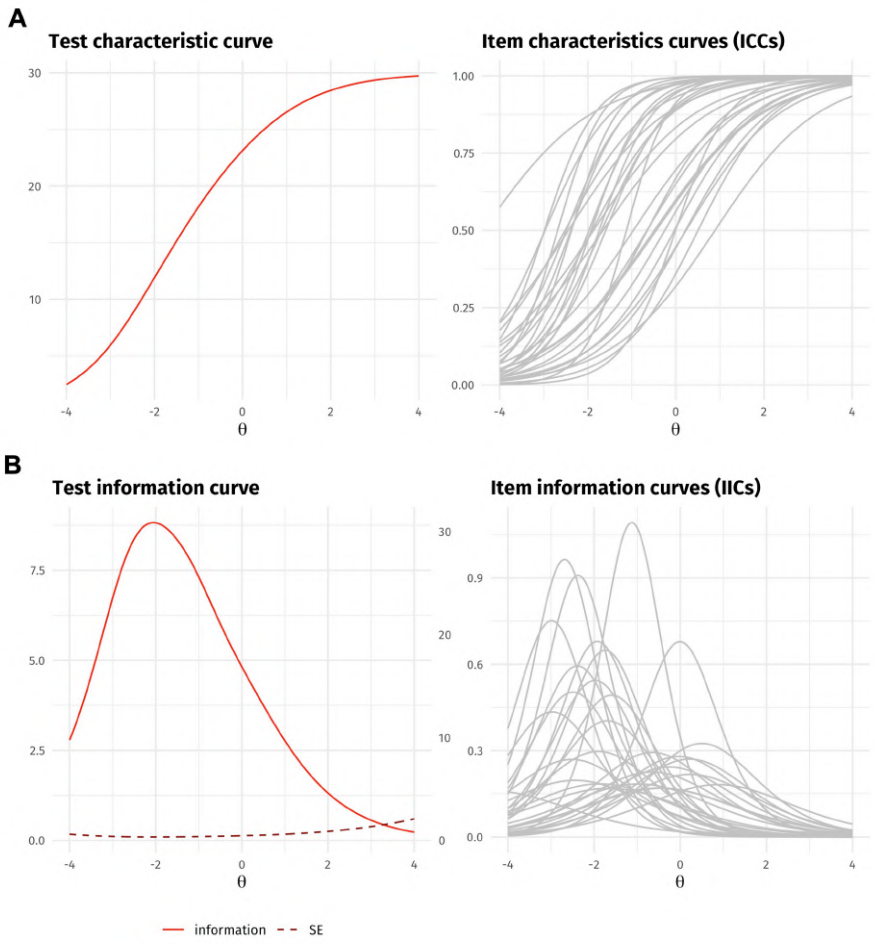


Figure 1: Characteristic Curves (A) and Information Curves (B) of the 30-item inventory under the two-parameter logistic model. The ICCs indicate high discrimination of most items particularly in the low-ability space; the TIC suggests that the scale is less informative for high-ability respondents.

Figure 1A (left panel) displays the test characteristic curve, illustrating the expected raw score on the 30-item inventory across the range of scores. This s-shaped curve indicates that as an individual's ability increases, their expected raw score on the INSPIRE inventory also increases, reflecting a coherent progression of performance. The right panel of Figure 1A shows the item characteristic curves (ICCs) for each of the 30 retained items. Consistent with our item selection discrimination criterion, these curves generally exhibit steep slopes, indicating that the selected items effectively discriminate between respondents at different levels of scientific evidence literacy. The spread of these curves along the score range further suggests that the item bank covers a meaningful range of difficulties, from easier items differentiating among lower-ability respondents to more challenging items for higher-ability individuals.

The overall precision and utility of the INSPIRE inventory are summarized by the test information curve (TIC), presented in Figure 1B (left panel). The TIC represents the sum of infor-

mation contributed by all items at each point along the ability continuum. The curve showcases coverage across the range of ability. There is, however, a substantial peak around θ=-2, indicating that the INSPIRE inventory provides the most precise and reliable measurement for individuals at the lower levels of scientific evidence consumption literacy. In other words, the instrument is particularly effective at differentiating among individuals who are less skilled respondents, which may be a desirable feature for interventions or policy decisions aimed at identifying and assisting those most in need of improved literacy. Figure 1B (right panel) displays the item information curves (IICs) for each individual item, illustrating how each item contributes to the overall test information. The peaks of these IICs generally align with their respective item difficulties, and their heights are a function of item discrimination, reinforcing that the retained items are informative. A further disaggregated overview of individual item characteristic curves and information curves, as well as an assessment of item fit, is provided in Supplementary Appendix A.2. The overall marginal reliability of the scale is high ($\omega_t = 0.81$).

All-in-all, the results from the IRT analyses suggest that the INSPIRE inventory offers good measurement precision across a relevant range of scientific evidence consumption literacy, particularly for individuals in the average and and lower-than-average ability range. Practitioners and researchers, however, can selectively employ items from the bank with higher difficulty parameters to shift the peak of information towards higher ability levels if their specific use case requires greater precision among more proficient individuals (see Guidance for implementation of INSPIRE section). This targeted informativeness, combined with the broad coverage of item difficulties, ensures that the instrument is well-suited for assessing competence in policy-relevant contexts.

Validation

Building on the data collection described in the previous sections, we conducted multiple validation exercises across three distinct samples to assess the psychometric and substantive validity of the INSPIRE inventory. These include (1) the Prolific respondent pool, (2) a group of policy students and professionals participating in formal training in data science, and (3) a pre-election forecasting sample collected in Germany during the 2025 federal elections.

Each sample provided complementary data for validating the inventory from different angles. For the Prolific sample, we collected measures of self-assessed quantitative knowl-

	(*Prolific sample*) Correlation with self-reported knowledge in the following areas:					
	Math (arithmetic, linear algebra, calculus)	**Probability and statistics**	**Data analysis**	**Scientific methods**	**Cause-and-effect reasoning**	**Consumption of scientific evidence**
Full scale (θ)	0.16	0.21	0.19	0.28	0.26	0.28

	(*Policy student and professional sample*) Correlation with self-assessment regarding their ability to:							
	describe data necessary to answer policy questions	**choose methods to answer policy questions**	**use graphs or math representatic to describe data**	**convert raw data points and present them**	**understand technical language in scientific studies**	**come up with arguments on the basis of evidence**	**identify research designs more robust for specific questions**	**design studies to answer policy questions**
Full scale (θ)	0.61	0.61	0.4	0.45	0.53	0.34	0.37	0.25

Table 3: Pearson correlations between full-scale INSPIRE θ scores and participants' self-assessments.

edge, scientifically controversial beliefs, and responses to an applied reasoning task involving evaluations of the suitability of studies for evidence needs. For the policy student and professional educational-setting sample, participants provided self-assessments of their capacity to work with scientific data in policy settings—for example, evaluating appropriate study designs and identifying relevant evidence. Finally, the pre-election forecasting sample received two inventory items focusing on data and visual literacy. We use these responses to examine how data reasoning relates to participants' ability to interpret visual forecast information and make accurate probabilistic judgments about electoral outcomes.

Taken together, these validation exercises allow us to assess the inventory's construct validity, criterion validity, and applied predictive utility. The following sections describe each validation approach in more detail and present evidence of the scale's robustness and relevance across policy-adjacent contexts.

Convergent validity To assess the convergent validity of the INSPIRE instrument, we examined associations between participants' θ scores and both self-reported domain knowledge for the Prolific sample and ability to engage in activities relating to applied evidence and policy work in the data science training sample.

First, we explored correlations between θ scores and Prolific participants' self-assessed knowledge across six domains: mathematics, probability and statistics, data analysis, scientific methods, cause-and-effect reasoning, and consumption of scientific evidence. As shown in the top panel of Table 3, INSPIRE scores are positively correlated with self-rated knowledge across several domains, including mathematics, probability and statistics, scientific methods, and cause-and-effect reasoning. The strongest associations are found for

Crucially, participants were also asked to make a series of judgments regarding the outcome of the upcoming election. Specifically, they were asked to report the expected federal-level vote shares for the six parties represented in parliament (the so-called second vote share) and the expected first vote shares of the candidates of the respective parties in their district. Furthermore, they had to guess the probability (on a 0 to 100 percent scale) of two randomly selected of several downstream scenarios of the election, including whether: (1) the conservative parties CDU/CSU would become the biggest faction in parliament (true), (2) the party Die Linke would receive at least three district mandates (true), (3) the Social Democrats would receive more than 20% of the votes (false), (4) the parties FDP, Die Linke and BSW would altogether not enter parliament (false), (5) the Social Democrats would receive more votes than the Greens (true), and (6) whether the Greens would receive more votes than the AfD (false). Furthermore, respondents were presented with results from a fictitious poll of 3,000 randomly selected voters for four parties (Party A: 45%, Party B: 25%, Party C: 20%, Party D: 10%) and asked to estimate the probability (as a percentage) that Party A would receive the most votes. Given the clear majority for Party A, we encoded responses of 80% or higher as indicating a correct intuition about the election outcome. For the expected vote shares, we computed the mean absolute error using the final election results as benchmarks. For the scenario-related questions, we used 100 as a benchmark for events that ultimately materialized and 0 for those that did not and calculated the average deviation across the given scenarios. Importantly, all those outcomes can be seen as closely related to information from published polls available prior to the election and most of them in the "very likely" or "very unlikely" territory. Available pre-election polls matched the election outcome remarkably well, which means that both awareness of those polls and, particularly important in our context, the ability to make sense of them can be seen as critical for people's performance in those tasks.

We then investigated whether performance on the individual INSPIRE items, as well as the composite item score, predicted performance in the four judgment tasks. To that end, we estimated a series of linear models, one for each outcome and key predictor (16 models in total). Furthermore, in all models we adjusted for key demographic and political interest variables, including educational attainment, gender, age, and political interest. In order to adjust for knowledge about party performance in opinion polls, we also included a measure that captured the mean absolute difference between party performance in the polls at the time of the survey and the perceived performance as reported by the respondent. This turned out to be the most influential variable for prediction accuracy in all models.

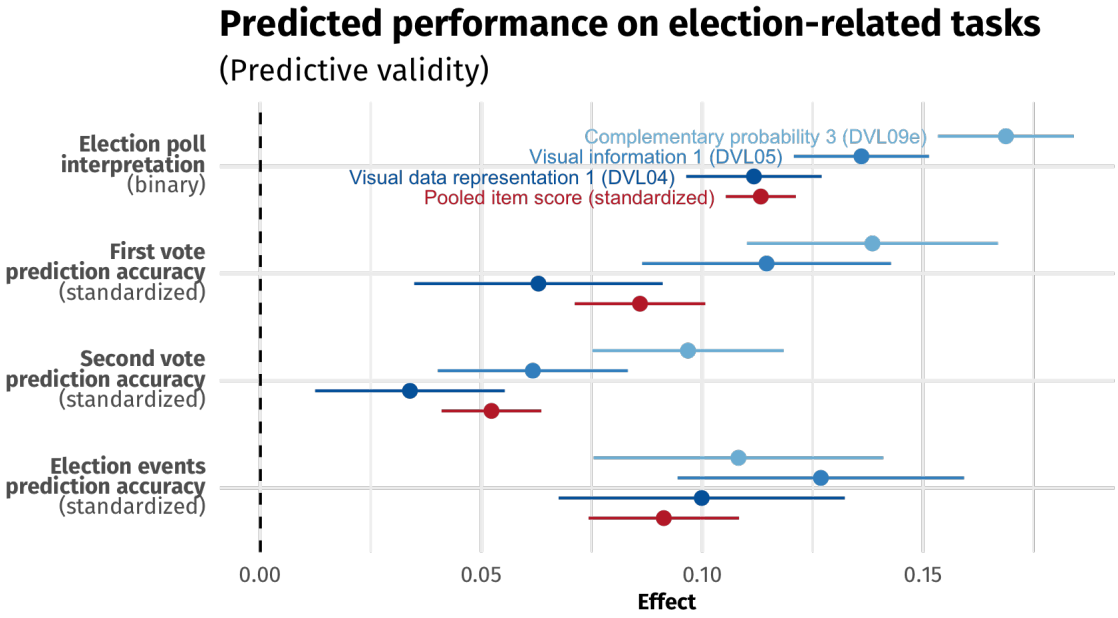


Figure 4: Effects of INSPIRE items on election prediction and judgment tasks. Effect estimates from linear regressions. Dots indicate point estimates, and horizontal lines show 95% confidence intervals. All models adjust for socio-demographic and knowledge-related controls. Sample sizes range from 15,302 to 15,570, depending on the specific outcome variable. Full model results are reported in Tables C4-C7.

Fig. 4 provides the marginal effects of all INSPIRE items as well as the composite score on respondents' performance in each of the four tasks. Across the board, data and visualization literacy scores prove to be robustly and strongly associated with respondents' judgment quality. This suggests that the visual and data literacy skills captured by INSPIRE offer predictive power beyond what is captured by general socio-demographic characteristics, political interest, and relevant knowledge, illustrating their relevance in forecasting performance. The findings from this validation exercise collectively support the predictive validity of the INSPIRE inventory, demonstrating its utility in assessing skills relevant to real-world probabilistic reasoning and forecasting accuracy.

Discussion

In this study, we introduced and validated the *Inventory for Numeracy, Statistics, and Policy-oriented Inference and REasoning* (INSPIRE), an instrument designed to assess scientific evidence consumption literacy within policy-relevant contexts. Our findings suggest that INSPIRE is a psychometrically robust and practically valuable tool for measuring this crucial set of skills, addressing a recognized gap in existing assessment measures by specifically targeting the unique demands of evidence consumption in public policy.

DVL9 (e). Complementary Probability 3
In a city, 40% of the population is vaccinated against a disease. If a person is selected at random, what is the probability that the person IS NOT vaccinated against the disease?
A. 0.2
B. 0.4
C. **0.6 ✓**
D. That is impossible to tell

Causal Reasoning (5 items)

CR8 (e). Observational Study
A researcher conducts a study comparing two groups: one that received an intervention and one that did not. The groups were not randomly assigned. What is a typical name for such a study design?
A. **Observational study ✓**
B. Longitudinal study
C. Experimental study
D. Case study

CR9 (e). Control Variables
When conducting a regression analysis in an impact evaluation study for a policy, what is the primary purpose of including control variables?
A. To increase the sample size
B. **To control for factors that could influence both the treatment and the outcome ✓**
C. To simplify the analysis
D. To ensure all data is normally distributed

CR10 (e). Causality and Policy Evaluation
What is a potential challenge in establishing cause and effect relations in policy evaluation?
A. Limited availability of data
B. Difficulty in identifying what would have happened without the policy
C. Common causes between the policy and the outcome
D. **All of the above ✓**

CR11 (e). Experimental Study
Which of the following study designs is best suited for determining the causal impact of a policy?
A. Observational study
B. Focus group study
C. **Experimental study ✓**
D. Case study

CR12 (e). Study Conclusions 2
Below you find a statement about evidence for a policy issue. What do you think: Does the evidence support the conclusion?

Evidence: The national rate of unemployment this year is lower than normal. Approximately 100,000 jobs were added to the workforce every month. Salaries experienced an average increase of 6%.
Conclusion: Unemployment is not a problem in the country.
A. **The conclusion is not supported by the evidence ✓**
B. The conclusion is supported by the evidence

CR13 (e). Study Conclusions 3
Below you find a statement about evidence for a policy issue. What do you think: Does the evidence support the conclusion?

Evidence: In a randomized control trial, some randomly selected parking booths in the city center were updated to accept card payments during a pilot phase. During this pilot, fines decreased by 20% in streets with the new system. Fines returned to regular levels after the pilot ended.
Conclusion: Enabling card payments in parking booths can increase compliance with parking fee payments.
A. **The conclusion is supported by the evidence ✓**
B. The conclusion is not supported by the evidence

Appendix B Scale construction

	Domain and sub-domain	Item code	Item label	Exclusion reason	Percent correct	Median time (s)	N
Scientific Literacy							
Scientific Literacy	Basic science	SEL04	Research credibility		0.98	14	175
Scientific Literacy	Basic science	SEL03	Quality of evidence		0.95	18	170
Scientific Literacy	Source reliability	SEL07	Trustworthy data		0.94	13	190
Scientific Literacy	Source reliability	SEL08	Trustworthy sources		0.94	24	172
Scientific Literacy	Scientific practice	SEL06	Scientific practice 2		0.9	21	176
Scientific Literacy	Basic science	SEL01	Scientific hypothesis		0.88	10	189
Scientific Literacy	Scientific practice	SEL05	Scientific practice 1		0.85	22	175
Scientific Literacy	Basic science	SEL02	Scientific consensus		0.5	9	176
Statistical Literacy							
Statistical Literacy	Machine learning models	SML05	ML and perpetuated bias		0.97	8	220
Statistical Literacy	Machine learning models	SML07	ML application		0.91	10	210
Statistical Literacy	Base rate	SML03	Base rate 1		0.88	39	160
Statistical Literacy	Machine learning models	SML04	ML accuracy		0.84	9	234
Statistical Literacy	Center and spread	SML01	Center and spread		0.74	36	163
Statistical Literacy	Center and spread	SML02	Sample sizes and uncertainty		0.65	36	162
Statistical Literacy	Machine learning models	SML06	ML and flawed predictions		0.6	29	193
Data Literacy							
Data Literacy	Visual information	DVL06	Visual information 2		0.95	24	261
Data Literacy	Numeracy	DVL01	Complementary probability 1		0.9	11	174
Data Literacy	Visual information	DVL05	Visual information 1		0.81	48	241
Data Literacy	Visual representation	DVL04	Visual data representation 1		0.57	33	235
Data Literacy	Numeracy	DVL02	Complementary probability 2		0.5	31	182
Data Literacy	Numeracy	DVL03	Percentages		0.48	42	181
Data Literacy	Problematic visuals	DVL07	Flawed visuals 1		0.41	58	253
Data Literacy	Problematic visuals	DVL08	Flawed visuals 2		0.37	54	245
Causal Reasoning							
Causal Reasoning	Experiments	CR04	Randomized control trials		0.96	10	160
Causal Reasoning	Causal reasoning	CR05	Confounder		0.85	24	196
Causal Reasoning	Valid causal conclusions	CR07	Study conclusions 1		0.82	26	181
Causal Reasoning	Correlation	CR03	Learning from correlation		0.69	27	192
Causal Reasoning	Causal policy	CR06	Causal effect of a policy		0.64	27	193
Causal Reasoning	Correlation	CR02	Correlation		0.6	20	197
Causal Reasoning	Correlation	CR01	Correlation vs. Causation		0.46	10	168
Scientific Literacy (excluded)							
Scientific Literacy	Basic science	SEL09 (e)	Peer review process	High Q3	0.97	10	178
Scientific Literacy	Basic science	SEL12 (e)	Evidence synthesis	Low Disc	0.9	10	176
Scientific Literacy	Scientific practice	SEL13 (e)	Scientific practice 3	Low Disc	0.88	17	180
Scientific Literacy	Source reliability	SEL15 (e)	Trustworthy science	High Q3	0.83	15	162
Scientific Literacy	Basic science	SEL10 (e)	Replicability	High Q3	0.8	12	202
Scientific Literacy	Scientific practice	SEL14 (e)	Scientific practice 4	Low Disc	0.58	19	159
Scientific Literacy	Basic science	SEL11 (e)	Peer reviews bias	Low Disc	0.32	7	173
Statistical Literacy (excluded)							
Statistical Literacy	Machine learning models	SML12 (e)	ML definition	Low Disc	0.96	8	227
Statistical Literacy	Center and spread	SML08 (e)	Mean calculation	High Q3	0.91	20	169
Statistical Literacy	Machine learning models	SML13 (e)	ML and overfitting	Low Disc	0.71	12	188
Statistical Literacy	Center and spread	SML09 (e)	Median calculation	High Q3	0.62	24	186
Statistical Literacy	Machine learning models	SML14 (e)	ML and supervised learning	Low Disc	0.53	19	219
Statistical Literacy	Base rate	SML11 (e)	Base rate 3	Low Disc	0.33	42	187
Statistical Literacy	Base rate	SML10 (e)	Base rate 2	Low Disc	0.32	47	186
Data Literacy (excluded)							
Data Literacy	Visual representation	DVL10 (e)	Visual data representation 2	High Q3	0.95	25	257
Data Literacy	Numeracy	DVL09 (e)	Complementary probability 3	High Q3	0.88	16	178
Causal Reasoning (excluded)							
Causal Reasoning	Causal reasoning	CR09 (e)	Control variables	High Q3	0.87	18	193
Causal Reasoning	Causal policy	CR10 (e)	Causality and policy evaluation	Low Disc	0.84	17	163
Causal Reasoning	Valid causal conclusions	CR13 (e)	Study conclusions 3	Low Disc	0.84	31	196
Causal Reasoning	Valid causal conclusions	CR12 (e)	Study conclusions 2	Low Disc	0.72	26	203
Causal Reasoning	Experiments	CR08 (e)	Observational study	Low Disc	0.43	24	169
Causal Reasoning	Causal policy	CR11 (e)	Experimental study	High Q3	0.3	15	204

Table B1: Overview of the items in the initial 52-item pool.

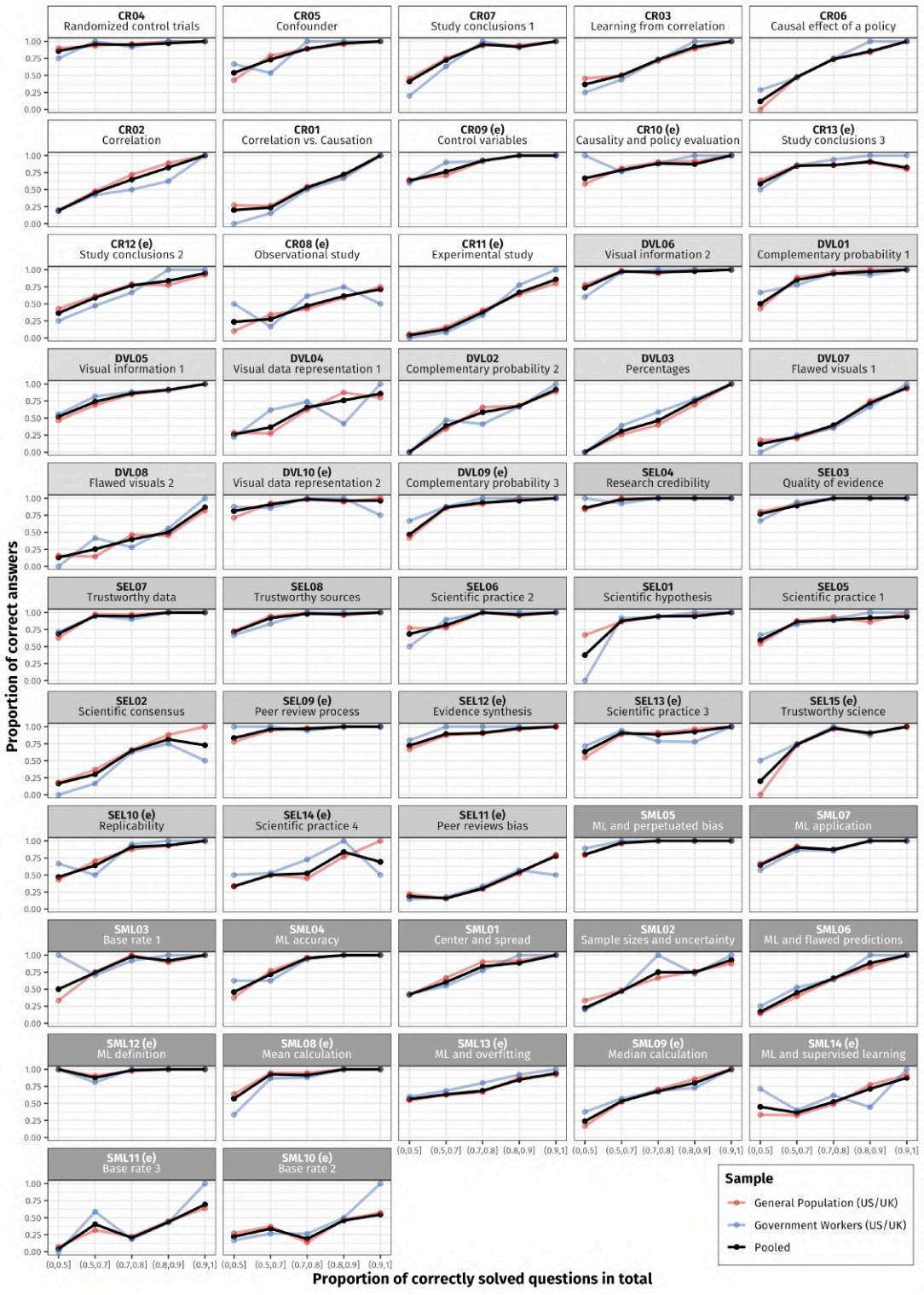


Figure B1: Nonparametric item characteristic curves by sample. This figure shows the proportion of respondents who answered each item correctly as a function of their overall test performance, segmented into bins of overall accuracy. Lines are plotted separately for general population respondents (red), government workers (blue), and the pooled sample (black). Each panel corresponds to a specific item, labeled by construct and item ID. These curves provide a nonparametric visualization of item characteristics, illustrating how the probability of correctly solving each item increases with overall ability.

	1PL	2PL		3PL		
Item	**diffi**	**discr**	**diffi**	**discr**	**diffi**	**guess**
SEL01	−2.37	1.61	−1.74	2.16	−1.26	0.27
SEL02	−0.04	1.06	−0.03	1.70	0.40	0.19
SEL03	−3.58	1.91	−2.37	4.83	−1.89	0.00
SEL04	−4.41	1.73	−2.99	6.58	−0.73	0.92
SEL05	−2.26	0.89	−2.44	1.20	−1.10	0.49
SEL06	−2.63	1.65	−1.92	1.89	−1.60	0.24
SEL07	−3.23	1.54	−2.39	1.51	−2.43	0.00
SEL08	−3.22	1.42	−2.49	1.35	−2.61	0.00
SML01	−1.45	0.82	−1.66	5.59	0.19	0.58
SML02	−0.63	0.95	−0.66	0.94	−0.65	0.00
SML03	−2.05	1.40	−1.59	2.11	−0.54	0.54
SML04	−2.00	1.27	−1.69	1.39	−1.57	0.00
SML05	−4.10	1.96	−2.68	16.17	−0.91	0.86
SML06	−0.39	0.82	−0.45	2.19	0.60	0.39
SML07	−2.60	1.04	−2.52	1.04	−2.41	0.08
DVL01	−2.61	1.47	−1.97	1.34	−2.10	0.00
DVL02	0.01	1.65	−0.01	1.62	−0.01	0.00
DVL03	0.22	0.93	0.23	2.13	0.81	0.26
DVL04	−0.24	0.99	−0.25	0.86	−0.28	0.00
DVL05	−1.70	0.86	−1.87	0.86	−1.86	0.00
DVL06	−3.63	1.32	−2.97	1.20	−3.18	0.01
DVL07	0.56	1.14	0.50	2.23	0.79	0.17
DVL08	0.80	0.85	0.89	1.85	1.21	0.19
CR01	0.24	1.04	0.22	1.88	0.56	0.17
CR02	−0.42	0.82	−0.49	1.18	0.29	0.28
CR03	−0.87	0.86	−0.97	15.04	0.39	0.49
CR04	−3.67	0.80	−4.38	2.43	0.09	0.92
CR05	−2.04	1.09	−1.90	1.18	−1.81	0.00
CR06	−0.69	1.08	−0.65	2.88	0.22	0.37
CR07	−1.69	2.09	−1.12	2.25	−1.10	0.00

Table B2: Model parameters from Rasch, 2PL, and 3PL models. The table presents item-level discrimination, difficulty, and guessing parameters, estimated using marginal maximum likelihood.

	AIC	SABIC	HQ	BIC	Log Likelihood	Chi-square	df	p-value
1PL	8,560.94	8,611.84	8,645.90	8,776.88	−4,228.47	—	—	—
2PL	8,529.81	8,631.62	8,699.73	8,961.70	−4,160.91	135.12	52	0.000

Table B3: Likelihood ratio test and information criteria for 1PL vs. 2PL models

	AIC	SABIC	HQ	BIC	Log Likelihood	Chi-square	df	p-value
1PL	8,560.94	8,611.84	8,645.90	8,776.88	−4,228.47	—	—	—
3PL	8,584.54	8,737.25	8,839.41	9,232.36	−4,136.27	184.40	104	0.000

Table B4: Likelihood ratio test and information criteria for 1PL vs. 3PL models

	AIC	SABIC	HQ	BIC	Log Likelihood	Chi-square	df	p-value
2PL	8,529.81	8,631.62	8,699.73	8,961.70	-4,160.91	—	—	—
3PL	8,584.54	8,737.25	8,839.41	9,232.36	-4,136.27	49.28	52	0.582

Table B5: Likelihood ratio test and information criteria for 2PL vs. 3PL models

Correlation matrix					
	Full scale (θ)	Science and evidence literacy (θ_{sel})	Statistical and ML literacy (θ_{sml})	Data literacy (θ_{dl})	Causal reasoning (θ_{cr})
Full scale (θ)	1.0				
Science and evidence literacy (θ_{sel})	0.5	1.0			
Statistical and ML literacy (θ_{sml})	0.6	0.2	1.0		
Data literacy (θ_{dl})	0.7	0.3	0.2	1.0	
Causal reasoning (θ_{cr})	0.6	0.2	0.2	0.2	1.0

Table B6: Correlation matrix of θ scores between entire INSPIRE set of items and domain-specific subsets.

Appendix C Auxiliary tables and figures

INSPIRE score			
Characteristic	**Beta**	**95% CI**	**p-value**
Education			
Doctorate degree (PhD/other)	—	—	
Graduate degree (MA/MSc/MPhil/other)	-0.07	-0.38, 0.23	0.6
Undergraduate degree (BA/BSc/other)	-0.19	-0.49, 0.12	0.2
Formal training in math, stats, data analysis			
None	—	—	
Minimal amount	0.08	-0.14, 0.31	0.5
Moderate amount	0.21	-0.03, 0.45	0.086
Significant amount	0.49	0.21, 0.77	<0.001
Age			
<30	—	—	
30-60	0.01	-0.15, 0.18	0.9
>60	-0.14	-0.52, 0.24	0.5
Sex			
Female	—	—	
Male	0.04	-0.10, 0.19	0.6
Prefer not to say	0.84	-0.05, 1.7	0.065
No. Obs.	470		
R^2	0.059		
Abbreviation: CI = Confidence Interval			

Table C1: Results from OLS regression. Predicting INSPIRe score as a function of respondent characteristics.

Choosing the study better suited to answer policy question.						
	Bivariate model			Model with controls		
Characteristic	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
INSPIRE score	2.52	1.89, 3.37	<0.001	2.58	1.93, 3.45	<0.001
Education						
Doctorate degree (PhD/other)				—	—	
Graduate degree (MA/MSc/MPhil/other)				0.52	0.17, 1.58	0.2
Undergraduate degree (BA/BSc/other)				0.50	0.17, 1.50	0.2
Age						
<30				—	—	
30-60				1.91	1.18, 3.08	0.008
>60				1.53	0.52, 4.49	0.4
Sex						
Female				—	—	
Male				0.73	0.47, 1.14	0.2
Prefer not to say				0.48	0.03, 8.81	0.6
N. observations	1,190			1,190		
Groups	470			470		
AIC	931			931		
BIC	946			977		
Abbreviations: CI = Confidence Interval, OR = Odds Ratio						

Table C2: Concurrent validity applied judgement task (Main text Fig 3 [left side]). Results from mixed-effects logistic regressions with respondent random effects

Rejecting innacurate scientific beliefs.						
	Bivariate model			Model with controls		
Characteristic	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
INSPIRE score	2.16	1.67, 2.80	<0.001	2.13	1.64, 2.75	<0.001
Education						
Doctorate degree (PhD/other)				—	—	
Graduate degree (MA/MSc/MPhil/other)				0.74	0.31, 1.79	0.5
Undergraduate degree (BA/BSc/other)				1.04	0.44, 2.50	>0.9
Age						
<30				—	—	
30-60				1.95	1.24, 3.06	0.004
>60				1.25	0.55, 2.83	0.6
Sex						
Female				—	—	
Male				1.41	0.93, 2.12	0.10
Prefer not to say				0.42	0.06, 3.12	0.4
N. observations	1,669			1,669		
Groups	250			250		
AIC	1,572			1,570		
BIC	1,588			1,619		
Abbreviations: CI = Confidence Interval, OR = Odds Ratio						

Table C3: Concurrent validity rejecting inaccurate beliefs task (Main text Fig 3 [right side]). Results from mixed-effects logistic regressions with respondent random effects