Artículo de revista:

Castro Torres, Andres F.; Akbaritabar, Aliakbar (2024) “The use of linear models in quantitative research”. Quantitative Science Studies (ISSN 2641-3337) https://doi.org/10.1162/qss_a_00294

RESEARCH ARTICLE

The use of linear models in quantitative research

Andres F. Castro Torres[1,2] and **Aliakbar Akbaritabar**[2]

[1]Center for Demographic Studies, Barcelona, Spain
[2]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany

Keywords: analysis framework, linear models, quantitative analysis, science-of-science

ABSTRACT

The diversity of analysis frameworks used in different fields of quantitative research is understudied. Using bibliometric data from the Web of Science (WoS), we conduct a large-scale and cross-disciplinary assessment of the proportion of articles that use linear models in comparison to other analysis frameworks from 1990 to 2022 and investigate the spatial and citation patterns. We found that, in absolute terms, linear models are widely used across all fields of science. In relative terms, three patterns suggest that linear model-based research is a dominant analysis framework in Social Sciences. First, almost two-thirds of research articles reporting a statistical analysis framework reported linear models. Second, research articles from underrepresented countries in the WoS data displayed the highest proportions of articles reporting linear models. Third, there was a citation premium to articles reporting linear models in terms of being cited at least once for the entire period, and for the average number of citations until 2012. The confluence of these patterns may not be beneficial to the Social Sciences, as it could marginalize theories incompatible with the linear models' framework. Our results have implications for quantitative research practices, including teaching and education of the next generations of scholars.

Uncorrected Proof

1. INTRODUCTION

Diversity is an asset for the practice of scientific research. The benefits of diversity apply to how phenomena are represented in concepts and theories and how data are collected and analyzed. The diversity of quantitative analysis frameworks used in different fields of science is under-studied (Koppman & Leahey, 2019; Leahey, 2005, 2008). A large-scale and cross-disciplinary assessment of the prevalence of linear models compared to other analysis frameworks is lacking. In the first part of this study, we measure the proportion of research articles that rely on linear models from 1990 to 2022. We refer to this measure as *linear model-based research* prevalence. We present separated results for six macro fields of science (OECD, 2007) and 254 subdisciplines (Milojević, 2020; Pudovkin & Garfield, 2002). In the second part, we study the spatial patterns of the country-level prevalence of linear model-based research in Social Sciences. Finally, we compare the average number of citations between articles using linear models and other statistical methods to assess the potential existence of a citation premium to linear model-based research, namely, a larger average number of citations to linear model-based research compared to research based on other methods.

While we remain neutral about the ideal prevalence of linear model-based research, investigating any sustained high prevalence, spatial patterns, and citation premiums related to it is essential. Assessing the implications of these findings can drive actions and discussions to

an open access journal

Check for updates

Citation: Castro Torres, A. F., & Akbaritabar, A. (2024). The use of linear models in quantitative research. *Quantitative Science Studies*. Advance publication. https://doi.org/10.1162/qss_a_00294

DOI: https://doi.org/10.1162/qss_a_00294

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00294

Supporting Information: https://doi.org/10.1162/qss_a_00294

Received: 26 April 2023
Accepted: 10 December 2023

Corresponding Author:
Aliakbar Akbaritabar
akbari.tabar@gmail.com

Handling Editor:
Vincent Larivière

Copyright: © 2024 Andres F. Castro Torres and Aliakbar Akbaritabar. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

preserve methodological pluralism in research, ensuring a robust and diverse approach to research.

1.1. What Are Linear Models?

Linear models are analytical frameworks in which the expected value of a dependent variable is modeled as a function of independent variables, as represented by Eq. 1.

$$E[\mathbf{Y}] = g(\mathbf{X} \times \mathbf{b}) \quad (1)$$

In Eq. 1, $E[.]$ stands for the expected value operator; $\mathbf{Y}$ represents a vector of length n containing the values of the dependent variable; $g(.)$ is the link function, $\mathbf{X}$ is an $(n \times p)$ dimensional matrix of p independent controls, mediators, or variables of interest; and $\mathbf{b}$ is a vector of length p that captures the marginal associations between the independent variables and $\mathbf{Y}$. This notation is a very general representation of generalized linear models where the expected value of the difference between fitted and observed values (i.e., errors or residuals) is zero. There are no further distributional assumptions (Dobson & Barnett, 2008).

Given data and under some assumptions regarding the distribution of $\mathbf{Y}$, the distribution and correlation of the errors, and the relationships among variables and across units of observations, the values of $\mathbf{b}$ (i.e., the regression coefficients) can be estimated (i.e., the marginal association between the independent variables and the outcome variable). When some variables in $\mathbf{X}$ are randomly assigned or exogenous, their coefficients could be interpreted in causal terms. Further, there are several techniques from Frequentist and Bayesian frameworks for assessing the statistical significance of all kinds of differences between regression coefficients and constant values (e.g., $b = 0$) and among them (e.g., $b1 > b2$). In addition, different combinations of independent variables can be evaluated and compared regarding their predictive accuracy concerning the outcome.

1.2. What Are Other Analysis Frameworks?

Alternative analysis frameworks (*other methods* herein) have been developed, for instance, in France in the 1960s and 1970s (Armatte, 2008; Lebaron & Le Roux, 2015) and in the United States more recently (Abbott, 2016; Koppman & Leahey, 2019; Ragin, 2014). These are all statistical techniques that do not require the specification of a dependent variable and do not aim to measure conditional associations or effects between the dependent and independent variables. Examples of other methods include network analysis, correspondence analysis, principal component analysis, sequence analysis, agent-based simulation, and cluster analysis.

Measuring the prevalence of linear model-based research in comparison to research using other methods is essential because the lack of diversity in methods, or *methodological monotheism*, as termed by Bourdieu and Wacquant (1992, pt. III), could translate into a lack of diversity in narratives, standardized questions, and lowering of the breadth of scientific research. While consolidating mainstream narratives indicates increased specialization (Leahey, Keith, & Crockett, 2010), it can also signal the marginalization of niche areas (Erola, Reimer et al., 2015; Moody & Light, 2006). As much as there are publication biases towards so-called null-result research (Fanelli, 2012), methodological monotheism (or the lack of pluralism) can potentially prevent or exclude atypical or disruptive research endeavors deemed not conform to the known traditions of standard science (Buyalskaya, Gallo, & Camerer, 2021; Lamont & Swidler, 2014; Leahey, 2008; Leahey, Lee, & Funk, 2021; Lin, Evans, & Wu, 2022; Teplitskiy, 2016; Uzzi, Mukherjee et al., 2013).

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

1.3. Are Linear Models Hegemonic in the Social Sciences Compared to Other Methods?

In the Supplementary material, we have assembled seven telling quotations from books and review articles from the 1970s to the present that illustrate scholars' concerns regarding the perceived high prevalence of linear model-based research as an analysis framework in the Social Sciences. This list includes Hirschman's assessment of fertility research, one of the most studied topics by the 1990s (Kohler, 2010; Mason, 1997):

> The standard social science model is that society works pretty much like a regression equation: the task is to find a right set of predictors, solve the equation, and discover what factors are the most important in predicting social outcomes. This framework does lead to empirical generalizations, but there seem to be endless qualifications about the measurement of variables, the meaning and interpretation of variables, the substitutability of one variable for another, and complex interactions with historical settings. If science is to discover parsimonious principles that explain complex patterns, we do not seem to be making progress (Hirschman, 1994, p. 256)

The question of whether linear model-based research is the dominant framework of analysis is therefore relevant for diversity, particularly for scientific fields where what constitutes evidence, explanation, and theory are socially contested and carry political ideologies and consequences, namely, the Social Sciences and the Humanities (Abend, 2008; Krause, 2016; Wallerstein, 1997). In the Social Sciences, higher levels of disagreement among scholars have been associated with the complexity of more significant problems (Lamers, Boyack et al., 2021). Controlled conditions for observing social phenomena are rarely achieved, and when they are, results may lack external validity (Lieberson & Lynn, 2002). Therefore, mutually exclusive theories coexist with their group of empirical testers building their case studies based on those theories (Schwemmer & Wieczorek, 2020; Watts, 2017; Watts & Dodds, 2011). Social issues are often processual, path-dependent, and synergetic and may be characterized by feedback effects between causes and consequences (Smith, 2013). There may also be feedback effects between describing a social phenomenon and the social phenomenon itself. These latter feedback effects arise from the inseparability of the subject and object of study (Bourdieu & Wacquant, 1992; Lieberson & Horwich, 2008).

In this context, our goal is to measure the prevalence of linear model-based research and its trend over the last three decades (1990–2022). We present results for six major fields of science to gain comparative insights. Next, we document global patterns of linear model-based research prevalence in the Social Sciences and discuss how they relate to existing inequalities in knowledge production and the Global South's intellectual dependency (Krause, 2016; Quijano, 2000). These discussions are based on existing critiques of the untested hegemony of linear model-based research (Gollac, 2004; Lebaron, 2003), its relations to the dominant analytical approach to studying the social world (Henrich, 2020; Nisbett & Masuda, 2003), and the Eurocentric historical nature of the Social Sciences (Bhambra & Holmwood, 2021; Zuberi & Bonilla-Silva, 2008).

2. DATA AND MEASURES

We use publications metadata from 40,603,923 articles from 1990 to 2022 indexed by Clarivate's Web of Science (WoS) provided by the German Competence Centre for Bibliometrics (German Competence Centre for Bibliometrics, 2021) via the Max Planck Digital Library. We excluded all research articles with no abstract. WoS is one of the most exhaustive bibliometric

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

databases, covering more than 21,000 journals (Clarivate, 2022; Visser, van Eck, & Waltman, 2021). English language journals from Western countries such as the United States and the United Kingdom are overrepresented in WoS. Together with China, these two countries comprise the largest shares of articles in the WoS database (Falagas, Pitsouni et al., 2008; Mongeon & Paul-Hus, 2016; Norris & Oppenheim, 2007). This bias implies that our results concern a specific area of existing research, mainly produced in the Global North and China (see Table 1 and Figure 1).

We limit our analysis to the Article document type, and we focus on abstracts to warrant a large temporal and spatial scope. Full-text analyses are impractical, and existing full-text databases for such analyses are more limited than bibliometric databases. We are aware that authors may not report the methods they use in the abstract. For example, if a method is standardized and widely used, there may be no need to mention it in the article's abstract (Blake, 2010; Cohen, Johnson et al., 2010; Westergaard, Stærfeldt et al., 2018). Different method-reporting patterns across disciplines and methods themselves may affect our results in ways we cannot predict in advance. However, it seems reasonable to assume that abstracts from theoretically oriented disciplines (e.g., the Humanities and the Natural Sciences) and abstracts from articles using mainstream methods are less likely to report them. In these hypothetical cases, the unreported methods are not central to the articles' contribution, and they are not a novelty that is worth highlighting to potential readers. Such types of biases have been documented regarding the country of study, whereby articles about the United States and Western European countries are less likely to report the country of study—the default location in Eurocentric terms—in titles and abstracts than articles about other regions of the world (Castro Torres & Alburez-Gutierrez, 2022; Kahalon, Klein et al., 2021).

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

2.1. Prevalence Measures

To partially counter the potential biases implied by looking only at abstracts, we compute four different measures for the prevalence of linear model-based research. These measures differ in the articles included in the denominator from a relatively small and specific set that allows us to approximate the prevalence of linear model-based research to a larger and more general analytical sample to check consistency over time.

Our primary measure of interest is the proportion of articles that report linear models in their abstract among all articles reporting any statistical method. In a second prevalence measure, articles that report linear models and other methods simultaneously are subtracted from the first linear model-based research prevalence. Third, we compute the prevalence of linear models using the total number of articles reporting methods or quantitative data as the denominator. The complete list of terms referring to statistical methods and quantitative data is available in the Supplementary material. Using these terms, we select 7,164,784 articles. The terms referring to statistical methods are organized according to whether they refer to a linear model (e.g., regression analysis) or other methods (e.g., correspondence analysis). Each term is classified as general or specific depending on whether they use a generic or specific name for describing the methods (e.g., regression vs. logistic regression). We built an initial list of 57 terms. We circulated this list among six colleagues with diverse disciplinary backgrounds, including health sciences, natural and social sciences, and statistics, asking them to add missing methods or comment on the ones already included. After a few exchanges and clarification with colleagues, we consolidated a list of 73 terms; this list was further extended to 163 while conducting the literature review and adding both hyphenated/nonhyphenated names and American and British spelling conventions. Of these 163 terms, 15 refer to quantitative

Table 1. Descriptive statistics for visibility measures and control variables, and articles' distribution according to the geographical location of the first author

Articles' visibility measures, control variables, and first authors' location	Articles reporting quantitative methods or data in their abstract				Articles reporting quantitative methods in their abstract			
	1990–2012		2013–2018		1990–2012		2013–2018	
	Value	S.D.	Value	S.D.	Value	S.D.	Value	S.D.
Average number of citations three years after publication	3.0	4.8	4.9	9.0	3.2	4.9	4.9	9.6
Proportion of articles with no citation three years after publication	0.28	0.45	0.17	0.38	0.26	0.44	0.16	0.37
Proportion of articles mentioning linear models in abstracts	0.34	0.48	0.37	0.48	0.64	0.48	0.67	0.47
Proportion of articles from single countries	0.64	0.48	0.55	0.50	0.63	0.48	0.54	0.50
Articles' year of publication	2005	5.8	2016	1.7	2005	5.8	2016	1.7

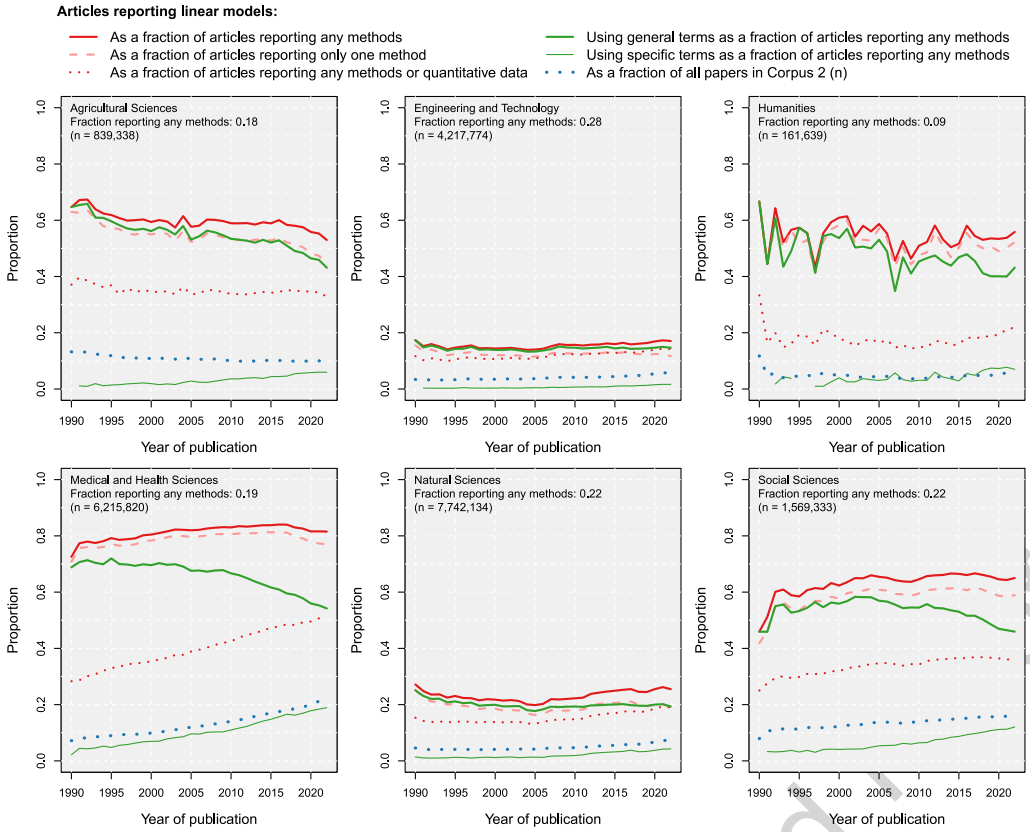
First author location	Total	%	Total	%	Total	%	Total	%
United States	167,737	42.8	116,243	30.5	89,728	42.4	64,883	30.5
United Kingdom	35,997	9.2	29,988	7.9	17,263	8.1	15,387	7.2
China	11,565	3.0	24,704	6.5	6,985	3.3	15,293	7.2
Germany	14,519	3.7	18,554	4.9	8,250	3.9	10,784	5.1
Australia	14,692	3.7	17,944	4.7	7,208	3.4	8,892	4.2
Canada	20,435	5.2	15,986	4.2	11,564	5.5	9,157	4.3
Spain	9,208	2.3	12,492	3.3	5,704	2.7	7,450	3.5
Netherlands	13,399	3.4	12,609	3.3	7,743	3.7	6,965	3.3
Italy	6,930	1.8	9,954	2.6	4,147	2.0	6,195	2.9
France	7,494	1.9	8,125	2.1	4,253	2.0	4,737	2.2
Central and Southern Asia	4,238	1.1	6,774	1.8	2,345	1.1	3,972	1.9
Eastern and Southeastern Asia	22,867	5.8	26,067	6.8	13,699	6.5	14,885	7.0
Europe and North America	41,637	10.6	52,028	13.7	21,974	10.4	28,132	13.2
Latin America and the Caribbean	6,218	1.6	9,821	2.6	3,390	1.6	5,620	2.6
Northern Africa and Western Asia	9,975	2.5	12,301	3.2	5,738	2.7	6,824	3.2
Sub-Saharan Africa	5,001	1.3	7,228	1.9	1,857	0.9	3,511	1.7
Total	**391,912**	**100**	**380,818**	**100**	**211,848**	**100**	**212,687**	**100**

data, 110 refer to linear models (48 general, 50 specifics, 12 causality-related words), and 38 refer to other methods (18 general, 20 specific).

Finally, we compute the prevalence of linear model-based research among all articles where abstracts include *one of* the following seven words: "model," "data," "evidence," "empirical," "results," "method," or "methods" **and** *one of* the following five words: "analysis,"

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024



Q3 **Figure 1.** Sustained high prevalence of linear model-based research in four of six macro fields of science. Time trends in four measures of the prevalence of linear model-based research (red lines) and relative distribution of general (thicker green solid line) and specific (thinner green solid line) terms for reporting linear models in articles' abstracts across the six OECD fields of science, 1990 to 2022.

"analyze," "analyse," "study," or "investigate" used once or more in the same abstract. We argue that these keywords allow us to identify potentially empirical papers dealing with processing, analyzing, or interpreting data. Using these terms, we enlarge our analytical sample to 13,720,556 articles. By enlarging the denominator, this last measure of prevalence helps us to assess the consistency of our results over time from a more conservative and strict perspective.

2.2. (In)visibility Measures

We compute two measures of papers' visibility: the total number of citations received in the three years following publication and whether the paper received at least one citation during the same period. The three-year window time warrants comparability between older and more recent papers, and it is justified because articles' citations mature in the third year after publication considering disciplinary differences (Bornmann & Tekles, 2019; Wang, 2013). Considering citations in only one year could penalize articles published in later months of the same year (Donner, 2018).

2.3. Measuring Linear Model Based Research Relative Visibility

To examine linear model-based research's advantage in visibility outcomes, we fitted a series of negative binomial (link function = log) and binomial (link function = logit) models predicting our two visibility measures. The negative binomial distribution is suitable for modeling the

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

number of citations in the first three years, given the strongly skewed distribution of this outcome. The binomial distribution is appropriate for modeling whether articles receive at least one citation. Our primary variable of interest in these models is whether or not the authors reported a linear model in their abstract. Hence, the regression coefficients of our variable of interest measure the difference in the average number of citations and the odds of having at least one citation between linear model-based research and quantitative research that report other methods or quantitative data.

2.4. Geographical and Temporal Differences

To account for geographical differences in citation patterns, we stratified our models by the top 10 countries regarding the number of articles and the rest of the countries grouped into the United Nations Sustainable Development regions (United Nations, 2017). To assess potential changes over time in linear model-based research visibility advantage, we ran separate models for articles published before and after the median year of publication (i.e., 2012). This partition favors the symmetry in the sample size and, therefore, the uncertainty of estimates in each sample. We control for the year of publication, and whether or not an article involves authors from different countries (i.e., it is a product of international collaboration). These control variables are essential because the number of citations and the proportion of papers without citations increases and decreases over time, respectively (Nielsen & Andersen, 2021; Seglen, 1992; Waltman, 2016). And articles involving authors from only one country receive, on average, fewer citations and are more likely to have zero citations (Gomez, Herman, & Parigi, 2022; Narin, Stevens, & Whitlow, 1991; Puuska, Muhonen, & Leino, 2014).

2.5. Summary of Visibility Measures and Control Variables

Table 1 shows descriptive statistics for the two visibility measures and the control variables along with articles' distribution across the top 10 countries and world regions. These figures are shown according to whether articles report quantitative data and/or methods in their abstracts and for two periods (i.e., 1990–2012 and 2013–2018).

Table 1 corroborates increasing citations over time in our sample according to the average number of citations and the proportion of articles with no citations three years after publication. Likewise, the proportion of articles reporting linear models displays relatively small increases over time in both samples (i.e., 0.34 to 0.37 and 0.64 to 0.67). As for control variables, Table 1 suggests an increased geographical diversity according to the location of the first author. Although the United States comprises a large fraction of articles in all samples, there is a 12 percentage point decrease across the two periods in both samples, from approximately 40% to 30%. Likewise, the fraction of articles involving only one country has decreased over time.

To categorize publications into fields of science (OECD, 2007), we use a mapping of OECD macro fields of science to WoS subject classifications. Because the field of science categories is not mutually exclusive, articles can be included in more than one category at a time. We allow this multiple counting for field-specific analysis (Figures 1 and 2). Instead, we consider unique articles (single counting) in our country-level analysis and statistical models, which only deal with social science publications. In addition, we include only different countries per article (e.g., if an article has multiple authors from the same country and one author from another country, we consider it a product of the two countries and use distinct country addresses per article to account for international versus single country publications).

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

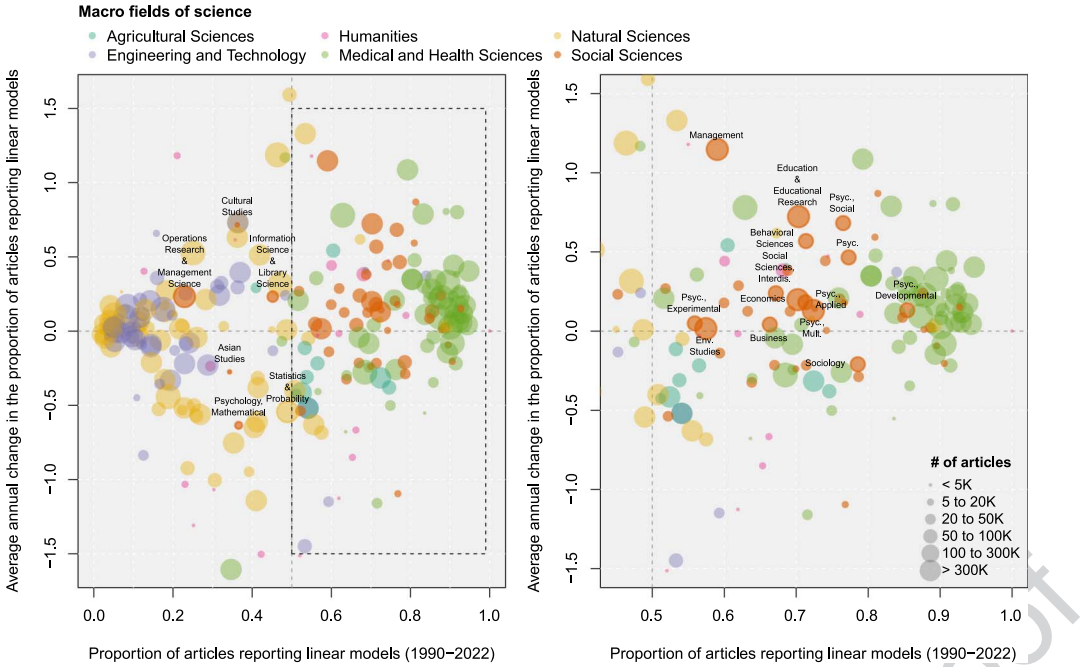


Figure 2. Social Sciences subdisciplines cluster above 50%, and the growing prevalence of linear model-based research. Distribution of subdisciplines according to the prevalence of linear model-based research and the average yearly change in the prevalence of linear model-based research as a fraction of articles reporting any methods from 1990 to 2022. Statistics & Probability and Social Sciences subdisciplines with more than 50,000 articles are labeled.

3. RESULTS

3.1. Trends in Linear Model-Based Research in Comparison to Other Methods Across Fields

Figure 1 shows time trends in the prevalence of linear model-based research for the six macro fields of science using our four different denominators (shades of red and blue) and the distinction between general and specific terms (green lines). This figure also shows aggregated numbers for each macro field, including the fraction of papers reporting methods, the fraction reporting methods or quantitative data, and the total number of articles.

The Humanities display the lowest fraction of articles reporting any method, with only 9% doing so. For the other disciplines, these fractions range from 18% in Agricultural Sciences to 28% in Engineering and Technology. The similarity of the fraction of articles reporting methods between the Natural Sciences and the Social Sciences speaks to the large scope of our list of terms. Slightly more than one-fifth of articles in both macro fields are selected using this list, and more than one-quarter in the case of Engineering and Technology. These fractions increase significantly when words referring to quantitative data are included, particularly in the Social Sciences, which further supports our assumption and the overall measurement logic.

In absolute terms, more than 7,000 articles reported linear models per year in all fields except for Agricultural Sciences (3,300) and the Humanities (260). Medical and Health Sciences rank first with more than 35,000 linear model-based articles per year, followed by the Natural Sciences (>15,000), the Social Sciences (>9,000), and Engineering and Technology (>7,500). In relative terms, the sustained high prevalence of linear model-based research (thick red lines) suggests that these methods are prevalent in four fields: Agricultural Sciences, Social Sciences, and the Humanities. The two-fold significance of linear model-based research for the Social Q4

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

Sciences in absolute and relative terms speaks to the influential role of these analysis frameworks for quantitative research in the field.

The Agricultural Sciences, Humanities, and Social Sciences display above 50% prevalence of linear model-based research over the 31 years of observation among articles reporting any methods. This pattern is robust to excluding articles reporting linear models and other methods simultaneously, particularly in the Social Sciences, where the prevalence of linear model-based research hovers around 65% for both prevalence measures. These patterns mean that under strict and conservative estimates of linear model-based research prevalence, approximately two-thirds of the research in Social Sciences has been conducted using the linear models analysis framework. The proportion of articles using general terms in this field (thicker green solid line) suggests that the high prevalence of linear model-based research is driven mainly by articles reporting general terms instead of specific models. Despite a positive trend over time, the fraction of papers reporting specific models in the Social Sciences (e.g., hierarchical models) is slightly above 10% by the end of the study period. This pattern indicates that the high prevalence of linear model-based research in Social Sciences has been largely unaffected by the increasing specialization of methods.

The consistency of the temporal trend for the proportion of papers reporting linear models as a fraction of articles reporting quantitative data (dotted red line) and all articles presumably performing empirical analyses (dotted blue line) suggests that the high prevalence of linear model-based research among papers reporting any statistical framework in the Social Sciences is unlikely to be driven by the selection of articles according to our list of statistical methods. The lower prevalence of linear model-based research among articles mentioning any methods and quantitative data may be related to articles that use quantitative sources as secondary or supplementary data without performing a direct statistical analysis to mention it in the abstract. Mentioning the use of secondary data without mentioning a statistical analysis framework may explain the significant difference between the red solid and dotted lines in the Humanities.

The patterns observed for Agricultural Sciences, Humanities, Medical and Health Sciences, and Social Sciences contrast with the low prevalence of linear model-based research in Engineering and Technology and Natural Sciences. Recall that the fraction of papers reporting methods and quantitative data is very similar or higher in these three subfields compared to the Social Sciences, which somehow accounts for potential field-specific reporting methods practices. In Medical and Health Sciences, more than 80% of the articles reporting methods reported linear models. We interpret this pattern as indicative of the appropriateness of our measurement strategy because medical and health subdisciplines often deal with randomized control trials and clinical trial data and research questions that involve identifying causal relationships between medical treatments and individuals' health or policy interventions and populations' wellbeing (Mitra, Roy, & Small, 2022). Linear models are well suited for these types of standardized questions. Observed trends in Engineering and Technology and the Natural Sciences could be driven by what is considered "methodological pluralism" in literature (Lamont & Swidler, 2014), as these fields have an inductive approach to science and would use a more diverse multitude of observational and analytical methods than, for example, the Medical and Health Sciences.

Figure 2 shows the scatter plot between the prevalence of linear models as a fraction of papers reporting any methods and the annual average percentage point change in the prevalence of linear models from 1990 to 2022 for most subdisciplines across macro fields of science. The right panel zooms into the areas of the plot where most of the Social Sciences subdisciplines cluster. According to this figure, Social Sciences subdisciplines, such as

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

Management, Educational Research, Psychology, and Economics, cluster in the top right areas of the plot, meaning above the 50% and growing prevalence of linear model-based research from 1990 to 2022.

The clustering in high and growing prevalences of linear model-based research among Social Sciences subdisciplines is only comparable to that of the Medical and Health Science. However, the latter displays a much higher prevalence of linear model-based research and relatively lower yearly average increase due to ceiling effects. As shown in the right panel of Figure 2, 42 of the 47 Social Sciences subdisciplines, including several with more than 50,000 indexed articles (labeled), display a prevalence of linear model-based research above 50%; 34 of them display positive temporal trends, meaning that linear model-based research has grown over the past 31 years. For example, Education & Educational Research displays a linear model prevalence of 70% along with a 0.72 average percentage point yearly increase from 1990 to 2022 (i.e., more than 22 percentage point absolute increase in the prevalence of linear model-based research during the analysis period).

This panel also shows that only five of the 47 subdisciplines in the Social Sciences display linear model prevalences below 50%: Asian Studies; Cultural Studies; Information Science & Library Science; Operation Research & Management Sciences; Psychology; and Mathematics. What is specific about these five subdisciplines is a question for future research.

3.2. The Center and Periphery Patterns in Prevalence of Linear Model-Based Research in the Social Sciences

Figure 3 shows the distribution of Social Sciences articles across countries in the log-scale in the top panel and the prevalence of linear model-based research at the country level at the bottom. The joint interpretation of these panels sheds light on the potential mechanism driving the sustained high prevalence of linear model-based research and allows us to speculate on potential causes for these patterns due to well-documented geographic power relations in knowledge production (Bhambra & Holmwood, 2021; Castro Torres & Alburez-Gutierrez, 2022; Krause, 2016).

As seen in the top panel in Figure 3, the countries of the Global North (i.e., Northern America, West and Northern Europe, and Australia) and China contribute the largest share of papers in the quantitative Social Sciences indexed in WoS. The United States is the most significant contributor with exp(–0.96) = 38% of the total papers in our sample, followed by the United Kingdom (7.7%), China (5.7%), Germany (4.1%), and Canada (4.1%). The five remaining countries of the top 10 producers (Australia, Spain, the Netherlands, Italy, and France) contribute between 1.9% and 3.9%. Together, these top 10 producers account for almost three-fourths of all articles in our sample (74.7%). These global patterns suggest that our analytical sample is not biased regarding the geographical origin of articles, as it corresponds accurately to the highest-producing countries in all fields as reported by other studies (Castro Torres & Alburez-Gutierrez, 2022; Muthukrishna, Bell et al., 2020).

The country-level pattern of linear model-based research prevalence (bottom panel) is partially consistent with the center–periphery and global North–South separation of countries, with some exceptions. Center–periphery ties, where the center exerts dominance in setting up research agendas and research methods, are typically observed in the global dynamics of knowledge production (Adame, 2021; Boshoff, 2009; Habel, Eggermont et al., 2014; Haelewaters, Hofmann, & Romero-Olivares, 2021; Krause, 2016). For example, countries with the lowest share of papers in Sub-Saharan Africa (periphery) display the highest linear model-based research prevalence. The same applies to underrepresented countries such

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

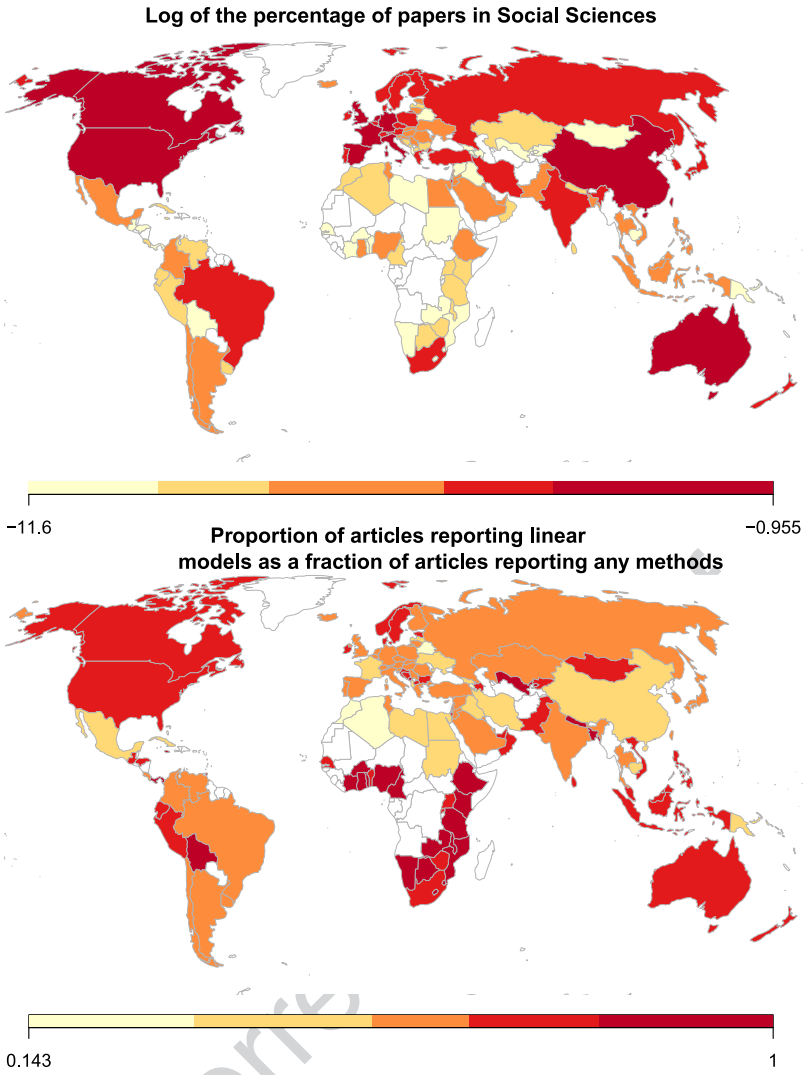


Figure 3. Global patterns of Social Sciences production and linear models prevalence, 1990–2021. Country-level share of articles in Social Sciences (top panel) and prevalence of linear model-based research for single-country articles (bottom panel). Information for countries with fewer than 10 articles in WoS is excluded. Cut-off points for the top panel are the 25th (0.006%), 50th (0.029%), 75th (0.35%), 90th (1.68%), and 100th (38.5%) percentiles. Cut-off points for the bottom panel (obtained via Jenks' algorithm) are 0.14, 0.33, 0.54, 0.65, 0.80, and 1.00. This selection of cut-off points assigned the top 10 countries on each scale to the first category and favors the readability of the remaining ones.

as Bolivia and Cuba in Latin America and the Caribbean, and Uzbekistan, Nepal, and Bangladesh in Asia. These countries are followed by the most prominent producers (i.e., the Global North), which, given their share of articles, drive the overarching trend of linear model-based research prevalence. The United States is the most prominent case, with 38% of the total papers and 74% reporting linear models among papers reporting any statistical framework. Australia and Scandinavian countries display similar levels of linear model-based research prevalence to that of the United States, yet none of them is as high as the prevalence observed in Sub-Saharan Africa and some countries in Asia and Latin America. France displays a distinctive pattern among the top 10 producers and Western

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

Europe: relatively large shares of Social Sciences articles (1.8%) and less than 50% prevalence of linear model-based research (49%). This geographic exception may be related to the development of the French School of statistics and its connection to the Social Sciences produced in this country's research institutions.

3.3. The Greater Visibility of Linear Model-Based Research Versus Other Statistical Analysis Frameworks

Figure 4 displays the association between reporting linear models and the odds of receiving at least one citation (left panel) and the average number of citations (right panel) within the three years after publication for the top 10 countries separately and for the remaining countries grouped in regions. If positive, these associations indicate visibility and citation premiums to linear model-based research, respectively. If negative, they indicate the opposite, namely, a visibility and citation penalty compared to research using other methods.

Receiving at least one citation in three years is a low-bar benchmark for articles' visibility. And the average number of citations within the three years after publications helps us distinguish articles that received more significant attention from the research community. The results are stratified for two periods: 1990–2012 and 2013–2018. The red and green markers show results for articles reporting any method and articles reporting any method or quantitative data, respectively. Our models control for whether the article is from a single country (vs. multiple countries) and the single year of publication.

The patterns in Figure 4 show partial citation premiums for linear model-based research in the centers of academic production (i.e., countries of the Global North). According to the left panel in Figure 4, from 1990 to 2012 (empty circles), there was a positive association between reporting linear models and receiving at least one citation, particularly for articles from the top 10 producer countries. Exceptions to this pattern include some top countries where the

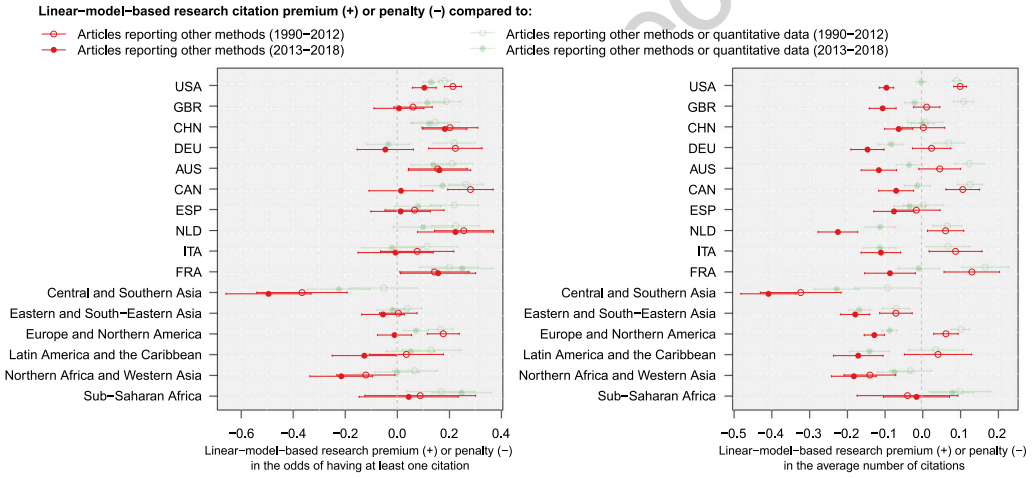


Figure 4. The visibility advantage of linear model-based research compared to quantitative research using other methods and quantitative data sources. Linear model-based research premium and penalty in the odds of having at least one citation (left panel) and the average number of citations above one (right panel) in the first three years after publication between articles reporting linear models and articles reporting other methods (red dots and lines) and articles reporting other methods or quantitative data (light green dots and lines). Regressions are stratified by countries with the 10 largest shares of articles in the Social Sciences (USA: United States, CHN: China, GBR: United Kingdom, DEU: Germany, NDL: the Netherlands, FRA: France, CAN: Canada, ITA: Italy, ESP: Spain, and AUS: Australia) and the UN-SDG regions for the other countries and by the median year of publication (i.e., 2012). Australia and New Zealand are included in Europe and Northern America, and other countries from Oceania are excluded due to the small sample size (1,350 articles).

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

association is positive, but the confidence interval contains zero (the United Kingdom, Spain, and Italy), and some regions where the association is negative (Central and Southern Asia, North Africa and Western Asia). Notably, the only positive and significant coefficient across regional groups pertains to Europe and Northern America, which speak to the differential value of analysis frameworks between the center and peripheral countries. From 2013 to 2018 (filled circles), this visibility premium was held in half of the top 10 producer countries: the United States, China, Australia, the Netherlands, and France, and weakened elsewhere; only Germany and Italy displayed negative coefficients for 2013–2018, and their confidence intervals include zero. Across regions, virtually all coefficients became negative (except for Sub-Saharan Africa), further highlighting the differential value of methods between top producers (center) and the rest of the world (periphery).

The average number of citations mirrors these patterns. According to the right panel in Figure 4, during the first two decades of analysis (1990–2012), articles reporting linear models received, on average, more citations than articles reporting other methods (red markers). This is true for all top 10 producers except China and Spain. For example, on average, articles from the United States reporting linear models had 10.7% more citations (coefficient = 0.1) than articles reporting other methods. These patterns of greater visibility for linear model-based research before 2012 are consistent when the reference group comprises articles reporting other methods or quantitative data (green markers). This consistency means linear model-based research potentially held a citation premium within quantitative research broadly defined. From 2013 to 2018, the citation premium observed in the right panel reversed among the top 10 producer countries and became more negative among regional groups. All the filled red dots lay on the negative side of the plot in the right panel, and only the confidence intervals for Sub-Saharan Africa include zero. These negative and statistically significant associations mean that during the last seven years of observation, linear model-based research was associated with a citation penalty, which signals a potential reversal in their prevalent use, at least regarding research visibility in both the center and the periphery.

To understand the potential drivers of this penalty, we replicated the right panel of Figure 4, excluding the top 10% and 5% articles in the number of citations within each country and region (see Figure S2 in the Supplementary material). These two replications show no citation penalty for any of the top 10 producer countries. The linear model-based research citation premium is observed for the United States, China, and Australia when the top 5% and 10% of cited articles are excluded from the sample. The citation penalty was held for all regions, although confidence intervals include zero for Europe, Northern America, and Sub-Saharan Africa. These patterns indicate that the reversal in the linear model-based research premium may be driven by a few articles (produced in the top 10 producer countries) using other methods and receiving a disproportionately large number of citations.

4. DISCUSSION

The prevalence of linear model-based research was above 50% for most Social Sciences subdisciplines from 1990 to 2022. This prevalence is higher in periphery countries compared to the centers of academic production. And linear model-based research displayed a citation premium for having at least one citation from 1990 to 2018, and from 1990 to 2012 for the average number of citations. The citation premium regarding the average number of citations reversed in the last years of analysis (2013–2018). Here we discuss these empirical results further by embedding them in previous discussions in the literature. We try to provide further context on the potential underlying causes of our results and elaborate on possible outcomes.

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

Linear models are very flexible as they can accommodate variables of different kinds (e.g., nominal, ordinal, interval, and continuous), and account for grouped observations in the dependent (e.g., repeated measures models) and independent variables (e.g., hierarchical or multilevel models). Linear models can be used with multiple link functions and probability distributions (Dobson & Barnett, 2008). These models have grown in complexity and specialization to accommodate better specific data types (e.g., categorical variables, duration, and sequence data) and to account for data features such as nested structures, serial correlations, and complex survey sampling (Cornwell, 2015; Courgeau & Leliévre, 1997; Mood, 2010). Additionally, linear models' outputs are easy to summarize, present, and interpret, notwithstanding issues of misuse and misinterpretations (Fanelli, 2012; Gelman & Loken, 2014; Leahey, 2005; Sterne, 2001; Vidgen & Yasseri, 2016; Yanai & Lercher, 2020). This flexibility makes them very appealing for a wide variety of scientific endeavors across fields of science, from data description to causal analysis.

However, it is hard to argue that linear models are suitable for answering all research questions within a field, for example, the Social Sciences (Buyalskaya et al., 2021; Grigoropoulou & Small, 2022; Leahey, 2005, 2008; Schwemmer & Wieczorek, 2020). According to previous research, the relative success of other methods in terms of the number of publications in the Social Sciences compared to linear models has been lower (Koppman & Leahey, 2019). Research has also shown that the more significant success of linear models' application in the Social Sciences cannot be fully explained by their flexibility and suitability (Leahey, 2005). Historical contingencies and power dynamics among individuals and institutions have played a role in the differential success of the linear model vs. other methods (Camic & Xie, 1994; Leahey, 2005; Lebaron, 2000; Porter, 1995). As a result, several contemporary assessments of social science research suggest that a potential hegemony of linear model-based research could marginalize specific questions and worldviews in disciplines such as economics, sociology, and demography (Lebaron, 2003; Sigle, 2021; Zuberi & Bonilla-Silva, 2008).

According to these studies, there are not enough discussions about the assumptions underlying the representation of the social world in the form of a linear equation (Abbott, 1988; Leahey, 2005). These representations rely on assumptions regarding the nature of the relations among variables (Abbott, 1988; Bourdieu, 1996; Zuberi & Bonilla-Silva, 2008), the imperative to "control for," standardize, or reduce data structure to identify pure effects (Gollac, 2004; Williams, 2019), and a particular understanding of causality as measurable only via quantitatively defined counterfactuals. These counterfactuals are typically obtained using technical procedures such as randomization, matching on observable characteristics, double differentiation, and instrumental variables (Smith, 2013).

These three essential features of linear-model-based research—centered on variables, concerned with pure effects or neat associations, and conceptualizing causal relations in terms of counterfactuals—align with a concrete understanding of theory as a set of testable propositions regarding the relations between outcomes and explanatory factors. According to Abend (2008), this is a valuable and legitimate definition of theory, which he terms "theory1," but it is not unique. There are at least six other uses of the word theory in Social Sciences, labeled by him as "theory2" through "theory7." Apart from Abend's "theory1," other theories are incompatible with the linear model framework. This lack of compatibility suggests that the hegemony of linear model-based research may marginalize specific theories, worldviews, and narratives about the social world, at least in quantitative Social Sciences.

Explicit critiques in this direction have been put forward since the late 1970s in different subdisciplines, including sociology, demography, economics, philosophy, and political sciences

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

(Abbott, 1988; Andler, Fagot-Largeault, & Saint-Sernin, 2002; Camic & Xie, 1994; Cornwell, 2015; Héran, 2006; Leahey, 2005; Lebaron, 2000; Porter, 1995; Sigle, 2021). Pierre Bourdieu, expressed it this way:

> The particular relations between a dependent variable (such as political opinion) and so-called independent variables such as sex, age and religion, or even educational level, income and occupation tend to mask the complete system of relationships which constitutes the true principle of the specific strength and form of the effects registered in any particular correlation. The most independent of 'independent' variables conceals a whole network of statistical relations which are present, implicitly, in its relationship with any given opinion or practice (Bourdieu, 1996, p. 103 [1979]).

Despite being widely cited in multiple subfields in the Social Sciences, Bourdieu's idea of the centrality of "the whole network of statistical relations" among variables and the methods he and his colleagues promoted to account for these relationships (e.g., multiple correspondence analysis), have not become part of the mainstream (Koppman & Leahey, 2019; Lebaron & Le Roux, 2015; Robson & Sanders, 2009). Instead, quantitative researchers have continued, primarily searching to isolate the effects of the independent variables on outcomes, particularly in the Anglophone social science tradition (Bry, Robette, & Roueff, 2016; Ollion, 2011).

5. CONCLUSION

The first part of our analysis validates our measurement approach. We are able to detect at least 20% of articles reporting methods in their abstracts except in Agricultural Science and the Humanities. Using these analytical samples we document substantial cross-field heterogeneity in the prevalence of linear model-based research. In absolute terms, the average number of articles reporting linear models ranged from a few hundred in the Humanities to more than 35,000 in Medical and Health Science. The Social Sciences ranked third, with almost 9,000 articles per year. In relative terms, the prevalence of linear model-based research ranged from low levels in Engineering and Technology (20%) to very high levels in Medical and Health Sciences (80%). The Social Sciences ranked second, with linear model-based research prevalences hovering around 65% for the entire analysis period.

The second part of our analysis suggests that in the Social Sciences, linear model-based research could be hegemonic. We conclude this based on three confluent patterns: the high, sustained, and growing prevalence of linear model-based research over time in Social Sciences in general and across its subdisciplines; geographical patterns in linear model-based research prevalence that are consistent with global inequalities in knowledge production; and the existence of a citation premium that has favored linear model-based research for the entire period in terms of having at least one citation (i.e., avoiding invisibility) and at least until 2012 for the average number of citations.

Despite being incapable of establishing causal relationships among the processes underlying these patterns, the salience of our results allows us to discuss the potential drivers and implications of these patterns. We rely on some previous, in-depth, single-discipline analyses (Koppman & Leahey, 2019; Leahey, 2005, 2008) and historical accounts of scholarly traditions' developments (Camic & Xie, 1994; Porter, 1995; Robson & Sanders, 2009, Chapter 2; Rouanet & Lépine, 1976) to inform our discussion and suggest future research areas.

The literature discussed the limitations of linear model approaches and the need to consider their presumptions (Abbott, 1988; Johnson-Hanks, Bachrach et al., 2011; Leahey, 2005; Sigle,

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

2021; Zuberi & Bonilla-Silva, 2008). These limitations would not be a concern if linear models were used as much as other methods. However, some discipline-specific studies have shown how linear modeling and hypothesis testing became normative in journals and fields such as economics, demography, sociology, psychology, and political sciences (Fanelli, 2012; Hirschman, 1994; Leahey, 2005; Lebaron, 2000; Ollion, 2011; Porter, 1995; Sigle, 2021). Our study extends these results to four macro fields of science and more in depth to 47 subdisciplines in contemporary quantitative Social Sciences by documenting a sustained, high, and growing prevalence of linear models-based research over the past 31 years. The presented prevalence suggests that we may be neglecting perspectives and approaches that do not conform to this analytical framework.

Studies on the use of quantitative methods in Sociology have examined the institutional and individual-level determinants of authors' use of alternative quantitative methods (Koppman & Leahey, 2019). According to these studies, conforming with mainstream analysis methods may be more beneficial for individual careers because existing institutional mechanisms reward this type of conformism. Authors with academic authority, seniority, prestige, and institutional credentials can afford the risk of exploring a nonmainstream perspective. And yet, their success in spreading the use of a given method is not always warranted, as shown by the case of correspondence analysis and qualitative comparative analysis (Koppman & Leahey, 2019). We cannot look at the institutional and individual-level determinants in all the 47 subdisciplines we study; however, the similarity among patterns between our results and previous in-depth single-discipline studies suggests that analogous mechanisms may be at play across these social science subdisciplines.

Linear models hegemony in terms of use, country-level distribution, and citation patterns could be consequential in several ways. First, hegemony could preclude the emergence, widespread use, and extension of potential path-breaking, atypical or disruptive approaches that could better solve current societal problems (e.g., rising social inequality, climate change, increased vulnerability of minorities). Second, global inequalities in knowledge production may be reinforced by the lack of methodological pluralism, as the capacity to develop new methods and bring them to the forefront of research is not evenly distributed across countries and institutions. Just as individuals with more extraordinary credentials and prestige can afford the risk of path-breaking analysis, countries, and regions in privileged positions (Zuckerman, 1970) (e.g., highly funded institutions) can lead research under alternative (risky) approaches (Guetzkow et al., 2004; Hofstra, Kulkarni et al., 2020; Lamont & Swidler, 2014; Laudel & Gläser, 2014; Luukkonen, 2012) without strategic concerns for publication or research evaluation (Akbaritabar, Bravo, & Squazzoni, 2021; Rijcke, Wouters et al., 2016). Third, to the extent that methodological pluralism is not promoted in research training programs and institutions, it may take some decades to appear, because, although the visibility premium has reversed, many generations of researchers and instructors were trained under a context that privileges linear model-based research.

Our study carries several limitations besides the relatively short temporal scope of our data from 1990 to 2022. First of all, we assume that methods reported in abstracts are the papers' analysis framework. Whereas this is likely the case for most empirical studies, our assumption may be incorrect if articles refer to methods for other reasons. Similarly, we are missing articles that do not report methods in their abstract, which directly affects our prevalence measures.

Based on research on reporting study countries in titles and abstracts, we find it reasonable to assume that mainstream methods (as mainstream study countries) are less likely to be reported in abstracts due to their widespread use. Research has shown that when researchers

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

study the so-called default cases, such as the United States, they are less likely to report the country of study in titles and abstracts (Kahalon et al., 2021; Castro Torres & Alburez-Gutierrez, 2022). We assume the same could apply to linear model-based research as it seems to be the default analysis framework according to cited literature. If this assumption holds, our analysis underestimates the prevalence of linear model-based research and overestimates the use of other methods. To address this limitation at least partially and in the case of a scientific field more familiar to authors (i.e., demography), we used full text information from 719 Open Access research articles from 2011 to 2022 published in *Demography*, the flagship journal of the Population Association of America. We find some evidence that favors our assumption. We limit these full text articles to only documents with above 20 paragraphs and exclude comments, notes, and letters, leading to the inclusion of 708 articles. We found that 91% of these papers mention linear models in their full text, vs. 47% that mention other methods. Further, we computed the weighted proportion of papers that report methods in their abstracts. We use the number of times that methods are mentioned in the full texts as the weights and focus on papers that mention methods at least five times (626 papers). These proportions were 36.6% for linear models and 41.2% for other methods, which indicates that, at least in *Demography* publications, as one illustrative example of quantitative social science research, linear models are less likely to be reported than other methods. Our weighting and sample selection strategies give more importance to papers that mention methods several times, which likely correlates with the substantive relevance of the method for the paper. Future research using full text from larger samples should test the validity of this assumption for quantitative research in other disciplines.

Continuing with limitations, we cannot evaluate critical aspects such as potential bias in the methods taught in graduate training programs, editors' and reviewers' potential preferences for certain types of analysis over others, or the concrete reasons for which researchers across different contexts decide to use a given method and influence others decisions (Lane, Teplitskiy et al., 2022; Merton, 1968). However, the magnitude of our results and the high prevalence of general terms for referring to analysis frameworks suggest that our results are unlikely to be wrong.

Finally, there are more comprehensive text analysis methods to identify all noun-phrase clauses or word combinations that our list might have overlooked (e.g., by selecting an anchor term and finding adjectives pairing up with it), as well as full-text analysis that could yield more comprehensive analytical samples of articles and statistical methods. Nevertheless, we decided on the most straightforward and strict method of search, aiming to find the lower bound of the prevalent use of analytical frameworks. Using more complex text analysis methods and full-text databases will likely increase the observed trends in the prevalence of linear model-based research and we encourage future research endeavors in this direction.

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

ACKNOWLEDGMENTS

We thank our colleagues Diego Alburez-Gutierrez, Enrique Acosta, Maarten Jacob Bijlsma, Beatriz Sofia Gil, Esther Dorothea Denecke, and Robert Gordon Rinderknecht, who reviewed and complemented the list of terms we have used here. We would also like to thank Diego Alburez-Gutierrez and Misha Teplitskiy for their helpful comments on the draft of our manuscript. We are thankful to the Nordic Network for the Science of Science for providing us with funding to present and receive feedback on our work at the Diversity in Science workshop at the University of Helsinki. We thank Monica Alexander for publicly sharing her workshop materials analyzing the full text of Open Access publications in *Demography* that we used for a robustness analysis comparing abstracts with full text.

AUTHOR CONTRIBUTIONS

Andres F. Castro Torres: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Aliakbar Akbaritabar: Conceptualization, Data curation, Investigation, Software, Validation, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This study has received access to the bibliometric data through the project "Kompetenzzentrum Bibliometrie" and we acknowledge their funder Bundesministerium für Bildung und Forschung (grant number 16WIK2101A). AF acknowledges funding from the Spanish Ministry of Science and Innovation through the Juan de la Cierva fellowship (FJC2020-042965-I) and the Nordic Network for the Science of Science.

DATA AVAILABILITY

We use data from the German Competence Centre for Bibliometrics (grant number 16WIK2101A). Restrictions apply to the availability of these data, which were used under license for the current study and are not publicly available. The replication data and scripts necessary to recreate our analysis and results are available upon request from the corresponding author.

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

REFERENCES

Abbott, A. (1988). Transcending general linear. *Sociological Theory*, *6*(2), 169–186. https://doi.org/10.2307/202114

Abbott, A. (2016). *Processual sociology*. Chicago IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226336763.001.0001

Abend, G. (2008). The Meaning of "Theory." *Sociological Theory*, *26*(2), 173–199. https://doi.org/10.1111/j.1467-9558.2008.00324.x

Adame, F. (2021). Meaningful collaborations can end "helicopter research." *Nature*. https://doi.org/10.1038/d41586-021-01795-1, PubMed: 34188244

Akbaritabar, A., Bravo, G., & Squazzoni, F. (2021). The impact of a national research assessment on the publications of sociologists in Italy. *Science and Public Policy*, *48*(5), 662–678. https://doi.org/10.1093/scipol/scab013

Andler, D., Fagot-Largeault, A., & Saint-Sernin, B. (2002). *Philosophie des sciences*. Gallimard.

Armatte, M. (2008). Histoire et Préhistoire de l'Analyse des données par J. P. Benzecri: Un cas de généalogie rétrospective. *Journal Electronique d'Histoire Des Probabilités et de La Statistique*, *4*(2), 1–24.

Bhambra, G. K., & Holmwood, J. (2021). *Colonialism and modern social theory*. Cambridge: Polity.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, *43*(2), 173–189. https://doi.org/10.1016/j.jbi.2009.11.001, PubMed: 19900574

Bornmann, L., & Tekles, A. (2019). Disruption index depends on length of citation window. *Profesional De La Información*, *28*(2). https://doi.org/10.3145/epi.2019.mar.07

Boshoff, N. (2009). Neo-colonialism and research collaboration in Central Africa. *Scientometrics*, *81*(2), 413–434. https://doi.org/10.1007/s11192-008-2211-8

Bourdieu, P. (1996). *Distinction: A social critique of the judgement of taste* (8th ed.). Cambridge, MA: Harvard University Press.

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology* (2nd ed.). Chicago, IL: University of Chicago Press.

Bry, X., Robette, N., & Roueff, O. (2016). A dialogue of the deaf in the statistical theater? Addressing structural effects within a geometric data analysis framework. *Quality and Quantity*, *50*(3), 1009–1020. https://doi.org/10.1007/s11135-015-0187-z

Buyalskaya, A., Gallo, M., & Camerer, C. F. (2021). The golden age of social science. *Proceedings of the National Academy of Sciences*, *118*(5), e2002923118. https://doi.org/10.1073/pnas.2002923118, PubMed: 33483415

Camic, C., & Xie, Y. (1994). The statistical turn in American social science: Columbia University, 1890 to 1915. *American Sociological Review*, *59*(5), 773–805. https://doi.org/10.2307/2096447

Castro Torres, A. F., & Alburez-Gutierrez, D. (2022). North and South: Naming practices and the hidden dimension of global disparities in knowledge production. *Proceedings of the National Academy of Sciences*, *119*(10), e2119373119. https://doi.org/10.1073/pnas.2119373119, PubMed: 35238625

Clarivate. (2022). *Web of Science Core Collection*. https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., & Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC*

Bioinformatics, *11*(1), 492. https://doi.org/10.1186/1471-2105-11-492, PubMed: 20920264
Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316212530
Courgeau, D., & Leliévre, E. (1997). Changing paradigm in demography. *Population*, *9*, 1–10.
Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). London: Chapman & Hall. https://doi.org/10.1201/9780367807849
Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics*, *12*(1), 330–343. https://doi.org/10.1016/j.joi.2018.01.012
Erola, J., Reimer, D., Räsänen, P., & Kropp, K. (2015). No crisis but methodological separatism: A comparative study of Finnish and Danish publication trends between 1990 and 2009. *Sociology*, *49*(2), 374–394. https://doi.org/10.1177/0038038514542495
Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF, PubMed: 17884971
Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. https://doi.org/10.1511/2014.111.460
German Competence Centre for Bibliometrics. (2021). *Kompetenzzentrum Bibliometrie (KB)*. https://bibliometrie.info/
Gollac, M. (2004). La rigueur et la rigolade. À propos de l'usage des méthodes quantitatives par Pierre Bourdieu. *Courrier Des Statistiques*, *1*(112), 29–36.
Gomez, C. J., Herman, A. C., & Parigi, P. (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, *6*(7), 919–929. https://doi.org/10.1038/s41562-022-01351-5, PubMed: 35637294
Grigoropoulou, N., & Small, M. L. (2022). The data revolution in social science needs qualitative research. *Nature Human Behaviour*, *6*(7), 904–906. https://doi.org/10.1038/s41562-022-01333-7, PubMed: 35347240
Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the humanities and the social sciences? *American Sociological Review*, *69*(2), 190–212. https://doi.org/10.1177/000312240406900203
Habel, J. C., Eggermont, H., Günter, S., Mulwa, R. K., Rieckmann, M., … Lens, L. (2014). Towards more equal footing in north–south biodiversity research: European and sub-Saharan viewpoints. *Biodiversity and Conservation*, *23*(12), 3143–3148. https://doi.org/10.1007/s10531-014-0761-z
Haelewaters, D., Hofmann, T. A., & Romero-Olivares, A. L. (2021). Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. *PLOS Computational Biology*, *17*(8), e1009277. https://doi.org/10.1371/journal.pcbi.1009277, PubMed: 34411090
Henrich, J. P. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Farrar, Straus and Giroux.
Héran, F. (2006). Toutes choses inégales d'ailleurs: Six études sur les discriminations. *Formation Emploi*, *94*, 5–10. https://doi.org/10.4000/formationemploi.2167
Hirschman, C. (1994). Why fertility changes. *Annual Review of Sociology*, *20*, 203–233. https://doi.org/10.1146/annurev.so.20.080194.001223, PubMed: 12318868
Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291. https://doi.org/10.1073/pnas.1915378117, PubMed: 32291335
Johnson-Hanks, J., Bachrach, C. A., Morgan, S. P., & Kohler, H.-P. (2011). *Understanding family change and variation: Toward a theory of conjunctural action*. Cham: Springer. https://doi.org/10.1007/978-94-007-1945-3
Kahalon, R., Klein, V., Ksenofontov, I., Ullrich, J., & Wright, S. C. (2021). Mentioning the sample's country in the article's title leads to bias in research evaluation. *Social Psychological and Personality Science*, *13*(2), 352–361. https://doi.org/10.1177/19485506211024036
Kohler, H.-P. (2010). Fertility trends and implications. In Z. Yi (Ed.), *UNESCO-EOLSS encyclopedia of life support systems: Demography*. EOLSS Publishers Co.
Koppman, S., & Leahey, E. (2019). Who moves to the methodological edge? Factors that encourage scientists to use unconventional methods. *Research Policy*, *48*(9), 103807. https://doi.org/10.1016/j.respol.2019.103807
Krause, M. (2016). "Western hegemony" in the social sciences: Fields and model systems. *The Sociological Review Monographs*, *64*(2), 194–211. https://doi.org/10.1002/2059-7932.12008
Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., … Murray, D. (2021). Investigating disagreement in the scientific literature. *ELife*, *10*, e72737. https://doi.org/10.7554/eLife.72737, PubMed: 34951588
Lamont, M., & Swidler, A. (2014). Methodological pluralism and the possibilities and limits of interviewing. *Qualitative Sociology*, *37*(2), 153–171. https://doi.org/10.1007/s11133-014-9274-z
Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., … Lakhani, K. R. (2022). Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Management Science*, *68*(6), 4478–4495. https://doi.org/10.1287/mnsc.2021.4107, PubMed: 36200060
Laudel, G., & Gläser, J. (2014). Beyond breakthrough research: Epistemic properties of research and their consequences for research funding. *Research Policy*, *43*(7), 1204–1216. https://doi.org/10.1016/j.respol.2014.02.006
Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, *84*(1), 1–24. https://doi.org/10.1353/sof.2005.0108
Leahey, E. (2008). Methodological memes and mores: Toward a sociology of social research. *Annual Review of Sociology*, *34*, 33–53. https://doi.org/10.1146/annurev.soc.34.040507.134731
Leahey, E., Keith, B., & Crockett, J. (2010). Specialization and promotion in an academic discipline. *Research in Social Stratification and Mobility*, *28*(2), 135–155. https://doi.org/10.1016/j.rssm.2009.12.001
Leahey, E., Lee, J., & Funk, R. J. (2021). What types of novelty are most disruptive? *Academy of Management Proceedings*, *2021*(1), 15612. https://doi.org/10.5465/AMBPP.2021.15612abstract
Lebaron, F. (2000). *La croyance économique—Les économistes entre science et politique*. Paris: Le Seuil.
Lebaron, F. (2003). Pierre Bourdieu: Economic models against economism. *Theory and Society*, *32*(5), 551–565. https://doi.org/10.1023/B:RYSO.0000004928.45684.d4

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

Lebaron, F., & Le Roux, B. (2015). *La méhodologie de Pierre Bourdieu en action: Espace culturel, espace social et analyse des données*. Paris: Dunod. https://doi.org/10.3917/dunod.lebar.2015.01

Lieberson, S., & Horwich, J. (2008). Implication Analysis: A pragmatic proposal for linking theory and data in the social sciences. *Sociological Methodology*, *38*(1), 1–100. https://doi.org/10.1111/j.1467-9531.2008.00199.x

Lieberson, S., & Lynn, F. B. (2002). Barking up the wrong branch: Scientific alternatives to the current model of sociological science. *Annual Review of Sociology*, *28*, 1–19. https://doi.org/10.1146/annurev.soc.28.110601.141122

Lin, Y., Evans, J. A., & Wu, L. (2022). New directions in science emerge from disconnection and discord. *Journal of Informetrics*, *16*(1), 101234. https://doi.org/10.1016/j.joi.2021.101234

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation*, *21*(1), 48–60. https://doi.org/10.1093/reseval/rvs001

Mason, K. O. (1997). Explaining fertility transitions. *Demography*, *34*(4), 443–454. https://doi.org/10.2307/3038299, PubMed: 9545624

Merton, R. K. (1968). The Matthew Effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56, PubMed: 5634379

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. https://doi.org/10.1162/qss_a_00014

Mitra, N., Roy, J., & Small, D. (2022). The future of causal inference. *American Journal of Epidemiology*, *191*(10), 1671–1676. https://doi.org/10.1093/aje/kwac108, PubMed: 35762132

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Moody, J., & Light, R. (2006). A view from above: The evolving sociological landscape. *The American Sociologist*, *37*(2), 67–86. https://doi.org/10.1007/s12108-006-1006-8

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., ... Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782, PubMed: 32437234

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, *21*(3), 313–323. https://doi.org/10.1007/BF02093973

Nielsen, M. W., & Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, *118*(7), e2012208118. https://doi.org/10.1073/pnas.2012208118, PubMed: 33558230

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences*, *100*(19), 11163–11170. https://doi.org/10.1073/pnas.1934527100, PubMed: 12960375

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. https://doi.org/10.1016/j.joi.2006.12.001

OECD. (2007). *Revised field of science and technology (FOS) classification in the Frascati Manual*. https://www.oecd.org/science/inno/38235147.pdf

Ollion, É. (2011). De la sociologie en Amérique. Éléments pour une sociologie de la sociologie étasunienne contemporaine. *Sociologie*, *2*(3), 277–294. https://doi.org/10.3917/socio.023.0277

Porter, T. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press. https://doi.org/10.1515/9781400821617

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113–1119. https://doi.org/10.1002/asi.10153

Puuska, H.-M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics*, *98*(2), 823–839. https://doi.org/10.1007/s11192-013-1181-7

Quijano, A. (2000). Coloniality of power and Eurocentrism in Latin America. *International Sociology*, *15*(2), 215–232. https://doi.org/10.1177/0268580900015002005

Ragin, C. C. (2014). *The comparative method: Moving beyond qualitative and quantitative strategies: With a new introduction*. University of California Press. https://doi.org/10.1525/9780520957350

Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—A literature review. *Research Evaluation*, *25*(2), 161–169. https://doi.org/10.1093/reseval/rvv038

Robson, K., & Sanders, C. (2009). *Quantifying theory: Pierre Bourdieu*. Cham: Springer. https://doi.org/10.1007/978-1-4020-9450-7

Rouanet, H., & Lépine, D. (1976). A propos de «l'Analyse des données» selon Benzécri. *Année Psychologique*, *76*(1), 133–144. https://doi.org/10.3406/psy.1976.28132

Schwemmer, C., & Wieczorek, O. (2020). The methodological divide of sociology: Evidence from two decades of journal publications. *Sociology*, *54*(1), 3–21. https://doi.org/10.1177/0038038519853146

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Sigle, W. (2021). Demography's theory and approach: (How) has the view from the margins changed? *Population Studies*, *75*(sup1), 235–251. https://doi.org/10.1080/00324728.2021.1984550, PubMed: 34902276

Smith, H. (2013). Research design: Toward a realistic role for causal analysis. In *Handbook of causal analysis for social research* (pp. 45–73). Cham: Springer. https://doi.org/10.1007/978-94-007-6094-3_4

Sterne, J. A. C. (2001). Sifting the evidence—What's wrong with significance tests? Another comment on the role of statistical methods. *BMJ*, *322*(7280), 226–231. https://doi.org/10.1136/bmj.322.7280.226, PubMed: 11159626

Teplitskiy, M. (2016). Frame search and re-search: How quantitative sociological articles change during peer review. *The American Sociologist*, *47*(2–3), 264–288. https://doi.org/10.1007/s12108-015-9288-3

United Nations. (2017). *United Nations sustainable development goals*. Regional groupings used in report and statistical annex. https://unstats.un.org/sdgs/indicators/regional-groups/

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157),

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024

468–472. https://doi.org/10.1126/science.1240474, PubMed: 24159044
Vidgen, B., & Yasseri, T. (2016). P-values: Misunderstood and misused. *Frontiers in Physics*, *4*, 6. https://doi.org/10.3389/fphy.2016.00006
Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112
Wallerstein, I. (1997). Eurocentrism and its avatars: The dilemmas of social science. *Sociological Bulletin*, *46*(1), 21–39. https://doi.org/10.1177/0038022919970102
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9
Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*, 0015. https://doi.org/10.1038/s41562-016-0015
Watts, D. J., & Dodds, P. (2011). Threshold models of social influence. In P. Bearman & P. Hedström (Eds.), *The Oxford handbook of analytical sociology*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199215362.013.20
Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Computational Biology*, *14*(2), e1005962. https://doi.org/10.1371/journal.pcbi.1005962, PubMed: 29447159
Williams, D. T. (2019). A call to focus on racial domination and oppression: A response to "racial and ethnic inequality in poverty and affluence, 1959–2015". *Population Research and Policy Review*, *38*(5), 655–663. https://doi.org/10.1007/s11113-019-09538-x
Yanai, I., & Lercher, M. (2020). A hypothesis is a liability. *Genome Biology*, *21*(1), 231. https://doi.org/10.1186/s13059-020-02133-w, PubMed: 32883349
Zuberi, T., & Bonilla-Silva, E. (2008). *White logic, white methods: Racism and methodology*. Lanham, MD: Rowman & Littlefield Publishers.
Zuckerman, H. (1970). Stratification in American science. *Sociological Inquiry*, *40*(2), 235–257. https://doi.org/10.1111/j.1475-682X.1970.tb01010.x

Uncorrected Proof

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00294/2351381/qss_a_00294.pdf by guest on 18 April 2024