**Mapping Methods in Contemporary Political Science Research:**

**An Analysis of Journal Publications (1998-2018)**

We describe here the process by which we compiled, coded, and analyzed the sample of 1,926 articles under study in “Mapping Methods in Political Science Research: An Analysis of Journal Publications (1998 – 2018).” We describe journal selection; article selection and sampling; code construction and codebook; coding procedure, and the procedures we followed to conduct our preliminary analysis. We also describe additional data collection about article authors and graduate methods syllabi. Our codebook is available upon request and will be posted to the Open Science Framework (OSF) website, associated with our Pre-Analysis Plan at the following URL: https://osf.io/uhma8/.

**Journal Selection**

In selecting journals to code, we compared the journal rankings of Giles and Garand (2007) to Garand et al. (2009). Giles and Garand’s (2007) journal rankings draw on citation-driven impact scores. Garand et al. (2009), by contrast, use a survey of political scientists in the United States, Canada, and the United Kingdom to compile their rankings. The top ten journals in Giles and Garand’s rankings and in the U.S. based rankings from Garand et al. overlapped substantially (once “sub-subfield” or topical journals such as *Legislative Studies Quarterly* and *Journal of Conflict Resolution* were excluded). The resulting list of journals included: *American Journal of Political Science* (*AJPS*), *American Political Science Review* (*APSR*), *British Journal of Political Science* (*BJPS*), *Comparative Political Studies* (*CPS*), *Comparative Politics* (*CP*), *International Organization* (*IO*), *International Studies Quarterly* (*ISQ*), *Journal of Politics* (*JoP*), *Perspectives on Politics* (*PoP*), *Political Research Quarterly* (*PRQ*), and *World Politics* (*WP*).

In order to narrow the list to ten journals to mirror the structure of Bennett et al.’s (2003) analysis and to place reasonable limits on the scale of our data collection effort, we eliminated *PRQ*, the “general-interest” journal with the lowest impact score. Of the remaining ten journals, five are general (*APSR, AJPS, BJPS, JoP,* and *PoP*), two predominantly focus on International Relations (*IO* and *ISQ*), two focus on Comparative Politics (*CP* and *CPS*), and one on a mix of International Relations and Comparative Politics (*WP*). While all of the discipline’s subfields are reflected in our journal selection, the empirical subfields are over-represented, as our study mainly focuses on the generation and analysis of data in empirical research.[[1]](#footnote-1) Also, the sample does not include a journal focused on American Politics because we posited that work in that subfield would be overrepresented in several general journals. Moreover, several highly-ranked American Politics journals are “sub-subfield” (i.e., they focus on a specialized slice of subfield topics, e.g., *Legislative Studies Quarterly* or *Politics, Groups, and Identities*), while our interest was in studying journals that represent all of the strands in one or more subfields. Nonetheless, in aggregate, the ten journals in our sample are more likely to publish research in Comparative Politics and International Relations than in Political Theory or American Politics.

**Article Selection and Sampling**

Once we had settled on the ten journals to include in our study, we developed a strategy for sampling articles. We decided to focus on work published between 1998 and 2018. That period overlaps with the period covered by the journal survey most similar to ours (Bennett, Barth, and Rutherford 2003). Moreover, examining that period ensured that we could assess the effect of the qualitative methods renaissance of the early 2000s, which we hypothesized catalyzed a proliferation of multimethod research articles in the short term. The exception to this time period is *PoP*, which the American Political Science Association (APSA) began to publish in 2003.

To ensure that our sample included comparable article types for each journal, we investigated the length and structure of work published over the relevant time period in the ten journals selected. Specifically, we looked through issues of each journal in three years (1998, 2008, and 2018) to identify submission types (e.g., research articles, replications, workshop proceedings, book reviews, and research notes). For each type, we noted a general description, whether peer review was necessary, the number of articles of each type that appeared in the issues analyzed, and the average word count for each type. We found most of this information using the respective journals’ submission guidelines. Reviewing the article types, we decided to only include in our sample peer-reviewed major research articles reflecting original research of (ostensibly) between 4,000 and 12,500 words in length (per article submission guidelines). We included work that met these criteria that was part of symposia and special issues, but excluded articles introducing or summarizing those article collections. We also excluded articles published in *APSR*’s centennial issue (vol. 100, no. 4), which also generally fell short of our word-length requirements. We excluded research notes, such as those published in *APSR*, to keep the sample homogeneous and the analysis comparable across articles.

In order to select our sample and to check the sample’s representativeness, we first sought to identify the population of articles that met our criteria. We listed each article that met our criteria in each issue (from 1998 to 2018) of each included journal in a Google Spreadsheet. For instance, if Volume 1 of Journal X had 4 articles, four corresponding entries were recorded using the scheme Journal.Year.Vol.Article No. This resulted in a population of 7,697 articles. In order to keep the study tractable but still analyze a significant number of articles, we elected to study 25 percent of the population (1,926 articles). Using Excel, we took a random sample of 2,126 entries, thereby including a surplus in our full sample in case we discovered in the course of the study that an article we sampled did not meet our selection criteria.[[2]](#footnote-2) We compared the journal and year distribution of the population to the journal and year distribution of our sample; the distributions showed very little difference. We recorded the full sample in a randomized list in a Google Spreadsheet. We then downloaded the relevant articles and stored them in a Box folder. Ultimately, we uploaded the first 1,926 articles to the Qualitative Data Analysis (QDA) software that we use for the project, Dedoose (see below).

**Code Construction and Practice Coding**

With the article sample in place, we developed and refined our codes through an iterative process. During the spring and summer of 2019 we drafted an initial set of codes that sought to address several core research questions: (1) what is the structure of authorship in articles published in highly ranked journals; (2) to what subfield do the articles correspond; (3) what techniques are employed to collect and/or generate the empirical base of those articles; and (4) what methods are employed to analyze the data that underpin that work? After generating an initial set of codes, we sought and integrated feedback from various faculty at our home institution and beyond.[[3]](#footnote-3) During the summers of 2019 and 2020, we continued to refine the codebook through 20 week-long rounds of practice coding. The initial author team (Hardy, Zachary Hunt, then a PhD student in the Department of Government at Georgetown, and Kapiszewski) — “practice coded” six sets of articles. In Spring 2020, Solomon replaced Hunt as a co-author. The current author team completed six rounds of practice coding; and Hardy and Solomon completed an additional eight rounds of practice coding.

In each round of practice coding, we coded 10 articles—five each from two journals. We identified the time period we wished each practice round to cover, and then articles from the selected journals were chosen at random. Coders used Dedoose QDA software to identify and highlight passages from each article that corresponded to the relevant codes. We choose Dedoose given the team’s experience and familiarity with the software, which attenuated start-up costs. To ensure that one coder’s application of the codes did not influence another’s, we used the “blind coding” setting in Dedoose. After each week of practice coding, we met (in person during Summer 2019 and remotely over Zoom during Summer 2020) to discuss coding discrepancies (i.e., why different codes were applied to the same passages or relevant codes not applied), weed out impractical codes, identify coding gaps, and make other refinements to the codebook to ensure its completeness and accuracy, and to facilitate precision and consistency across coders.

We finalized the codebook following practice round 14. The final codebook can be found in Appendix B of this pre-analysis plan.

Thereafter, we began to record measures of inter-coder reliability (ICR) between Hardy and Solomon as they continued to refine and align their coding. We also initiated twice-weekly practice rounds. To assess the consistency of Hardy’s and Solomon’s coding decisions, we used Cohen's kappa, a standard measure of ICR (Allen 2017a). Consistent with common interpretations of the Cohen’s kappa (Allen 2017a), we concluded practice coding after achieving a kappa coefficient of 0.80 or greater in two consecutive rounds of practice coding. To ensure that our assessment of ICR was robust to multiple measures, we also calculated the Krippendorff’s alpha (Allen 2017b) coefficient. The evaluations of consistency by the two coefficients were functionally equivalent across all six coding rounds in which we used them. We provide Cohen's kappa and Krippendorff’s alpha results for practice rounds 15 through 20 in Table 1A.

**Table 1A: ICR Scores for Summer 2020 Rounds of Practice Coding**

|  |  |  |
| --- | --- | --- |
| **Round** | **Cohen’s kappa** | **Krippendorff’s alpha** |
| 15 | 0.707 | 0.70702 |
| 16 | 0.721 | 0.72118 |
| 17 | 0.795 | 0.79478 |
| 18 | 0.769 | 0.769 |
| 19 | 0.843 | 0.84283 |
| 20 | 0.929 | 0.92876 |

**Formal Coding Procedure**

The coding procedure for each article was as follows. First, the coder identified an article to code from the full sample (described above). Second, the coder opened the article by searching for the article title in the Media tab in the Dedoose software. Third, the coder identified the article’s descriptive meta-data, which include both the journal in which the article appeared and the year in which it was published. The coder used the Descriptors tab in Dedoose to identify the appropriate journal-year pair, or the Create and Link Descriptor option to create a new pair if the appropriate pair was not available. Fourth, the coder used the Text Selection mode in Dedoose to highlight passages that corresponded to the codebook codes.

The coder analyzed the entire body of the article. Although there is some heterogeneity in the structure of published research in political science, some types of information were more likely to appear in specific article sections. In general, we drew subfield information from the abstract or introductory paragraphs of the article. For articles with an empirical basis, we used sections about research design, data, and methods—which often follow brief surveys or critiques of relevant academic scholarship—to identify the empirical information and methodological focus of the research. If the authors drew policy recommendations from their analysis, these statements typically appeared in the article’s concluding paragraphs. We included information in within-text appendices but excluded information in online supplemental materials, both for the sake of expediency and to hold article text constant over time.

We adopted both manual and computer-assisted strategies to minimize coder error and ensure consistency and precision during the hand-coding process. After completing the codes for an individual article, the coders double-checked that they had applied codes for Descriptors (Journal-Year), Authorship, Subfield, Empirical Base (unless a pure formal modeling article), and Method. This double-checking process also applied to articles that received codes for the source or type of human-participants data (Observational HP or Experimental HP); more than one code for empirical base (Tendency in Empirical Base); and more than one method code (Methodological Focus of MMR). As a computer-assisted backstop after completing each subset of approximately 500 articles, we used the statistical software R to confirm that each article had received these codes. For articles that had not received appropriate codes, the coder who originally coded the article reviewed and corrected the coding.

Additionally, we conducted weekly, and later bi-weekly, ICR checks to ensure that the coders’ decisions remained consistent. Each week (or every other week), we identified approximately three articles from each coder’s sample (a total of six articles) for the other coder to re-code. The coders used the blind-coding procedure discussed above to re-code each other’s articles. We calculated the Cohen’s kappa for the six articles. We provide the ICR results in Table 2A.

**Table 2A: ICR Scores for August 3, 2020 – August 28, 2020**

|  |  |
| --- | --- |
| **Week of** | **Cohen’s kappa** |
| 8/3/20 | 0.742 |
| 8/10/20 | 0.688 |
| 8/17/20 | 0.82 |
| 8/28/20 | 0.846 |
| 9/28/20 | 0.862 |
| 10/12/20 | 0.72 |
| 10/26/20 | 0.80 |
| 11/9/20 | 0.793 |
| 11/30/20 | 0.812 |
| 12/14/20 | 0.851 |
| 1/4/21 | 0.855 |
| 1/18/21 | 0.855 |
| 2/1/21 | 0.85 |
| 2/15/21 | 0.891 |
| 3/1/21 | 0.891 |
| 3/15/21 | 0.878 |
| 3/29/21 | 0.935 |
| 4/12/21 | 0.829 |
| 4/26/21 | 0.776 |
| 5/10/21 | 0.834 |
| 5/24/21 | 0.781 |
| 6/7/21 | 0.801 |
| 6/21/21 | 0.798 |

In some instances, lower Cohen’s kappa scores resulted from the cascading effects of a single coding mistake. For instance, a coder missing a human-participant based form of data generation in an article based on both pre-existing and author-generated data would result in two additional coding errors (failing to code the type of respondents and the tendency in the article’s empirical base). The overall average for all double-coded articles (N=134) was 0.818.

During the period of formal coding, the project team met weekly, and later bi-weekly, to discuss progress, ask questions, and consider the most significant inter-coder discrepancies. Coding was completed on approximately June 21, 2021.

A few additional notes round out our description of our coding procedure. Given the wide variety of scholarship in the sample, we sometimes had to employ external resources to learn enough about the empirical matter at hand (e.g., the profile of a group an author interviewed) in order to accurately and consistently apply our codes. We sought to rely on legitimate sources in conducting this research. Given the prevalence of Comparative Government and International Relations work in the sample, many articles dealt with non-English speaking contexts; we sometimes used Google translate to facilitate accurate coding.

As suggested above, codes were generally applied according to how the author described their data collection / generation and analysis; we typically did not assess the correspondence between how an author described their analytic process and how they carried out the analysis in the article. In some circumstances, these coding guidelines were a helpful check against inferential “overclaiming”: in some articles based on multivariate regression, for example, authors claimed to have identified a credible source of exogenous variation but fell short of causally identified research designs. Minimizing the degree to which we drew inferences from the author’s execution of their research design helped us to avoid bias in the data. In no instance did we assess the quality of any data collection, generation, or analysis process.

For the most part, our codes do not capture the steps of the research process that authors take between collecting / generating data, and analyzing those data to draw conclusions. For instance, we do not code the processes an author uses to organize / structure / reduce raw information collected, or ultimately to generate data from that information (beyond “text mining” and “discourse...analysis”). We decided not to attempt to code on these steps in the research process because they are the steps least likely to be clearly described in published scholarship.

**Additional Author-Level Data**

We also collected two categories of author-level data. First, we combined data about the self-described gender and racial identities of authors from the APSA membership database with our list of each article author. Second, we collected new data about authors’ PhD graduation year and training through a separate process that relied on information from authors’ personal or institutional websites. We describe both data cleaning and collection processes below.

*Finalizing the demographic data from APSA*

The APSA data include the self-described *gender* and *race / ethnicity* for each author. The gender variable includes categories for male, female, and non-binary, as well as a category for authors who decline to disclose their gender identity. The race / ethnicity variable includes a series of default racial / ethnic categories, which we describe in Section 3B below, as well as (1) an open-entry “Other” category and (2) a category for authors who decline to disclose their racial / ethnic identity.

The APSA data include (1) 1,293 “perfect matches,” which corresponded to a single entry in APSA’s membership database; (2) 321 “no matches,” for which there was no record; and (3) 1,129 “uncertain matches,” for which there were multiple entries in the APSA database that overlapped with our author database. To further clarify the uncertain matches, we calculated the string distance between the name of the author’s institutional affiliation in our database, and the name of the same in APSA’s. We also identified the entries for which our database identified an institutional affiliation but APSA’s did not, or vice versa. We automatically included in the final list entries with an “institutional name” string distance less than two (632 entries) or a missing institutional affiliation (121 entries). We flagged as “uncertain” all entries with a string distance of greater than two between our institution name and APSA’s.

This comparison yielded 376 remaining names, for which the names were equivalent but for which the string distance between the name of the author’s institution in our database and APSA’s institution name was greater than two. To reconcile these uncertain names, we compared by hand the name strings from our original database and from APSA’s database. We created a binary “match” variable, for which a value of 1 indicates that the names match, and a value of 0 indicates that they do not. We omitted all observations for which the value of the “match” variable was 0.

We used the following coding rules to manually compare the name strings in the two databases:

* If the two names had identical middle initials and identical names, we marked “match” as 1. If no middle initial was included and no other names matched, we marked “match” as 1. If middle initials were specified in both names and the middle initials did not match, we marked “match” as 0.
* If the author listed an institutional affiliation in the APSA database that was identical to the institutional affiliation in our database, but in a different language (e.g., Universiteit Maastricht = Maastricht University), spelled with different punctuation or an acronym (e.g., Indiana University Bloomington = Indiana University, Bloomington, or University of California, Los Angeles = UCLA, respectively), or referred to using a subunit or metonym (e.g., West Point University = United States Military Academy)---we marked “match” as 1.
* If, based on further research on the author’s personal or faculty website, including their CV, the institution with which the APSA database associated the author was a previous or more current affiliation than the affiliation listed in our database---we marked “match” as 1. If the university referenced in our database was different from the university in the APSA database and the author’s website or CV did not indicate that their institutional affiliation had changed, we marked “match” as 0.
* If, based on further research on the author’s personal website, faculty website, or CV, the author had changed their first name since they published their latest article in the database (e.g., Evgeny Finkel = Eugene Finkel)---we marked “match” as 1.

The APSA data also include information about the author’s self-described *subfield*.

*Collecting information about PhD graduation years*

In response to feedback that we received during a paper presentation at the 2021 APSA Annual Meeting, we also collected information about the PhD graduation years and disciplinary training of the authors in the sample. In Section 4 below, we describe our analytic procedure for testing hypotheses about the relationship between PhD graduation year and training---our measure of initial exposure to new analytic methods in the discipline---and methods of data collection and analysis.

We focused exclusively on PhD training in political science because we are interested in the relationship between methodological training in the discipline and methodological practices. For this reason, we did not collect information on PhD training in disciplines beyond political science and other cognates (e.g., politics, government).

Our analysis employs two main author-level variables. The first, *PhD year*, includes the following values:

1. **PhD Year**: An integer value recording the year in which the author received their PhD in a relevant program;
2. **PNDG**: The author is currently enrolled in a relevant PhD program;
3. **No PhD**:The author does not hold a PhD or the author holds a PhD in a discipline other than Political Science, Politics, or Government. To ensure that we capture disciplinary training in political science, we exclude the following common programs from the category of “cognate disciplines of political science”: (a) Public Administration; (b) Public Policy; (c) Political Communication; (d) Policy Analysis; (e) Social Science; and (f) International Development. We also exclude other social science disciplines such as Economics or Social Science;
4. **NAV**: The year the author received a PhD in a relevant discipline is either unavailable or it is unclear whether the author holds a PhD.

The second, *PhD institution*, includes the following values:

1. **PhD institution name**: The name of the institution that granted the author their relevant PhD or at which they are currently pursuing a relevant PhD. The PhD institution variable only carries this value if *PhD year* carries an integer value or PNDG.
2. **No PhD**:The author does not hold a PhD or that the author holds a PhD in a discipline other than Political Science, Politics, or Government. To ensure that we capture disciplinary training in political science, we exclude the following common programs from the category of “cognate disciplines of political science”: (a) Public Administration; (b) Public Policy; (c) Political Communication; (d) Policy Analysis; (e) Social Science; and (f) International Development. We also exclude other social science disciplines such as Economics or Social Science;
3. **NAV**: The institution that granted the author their relevant PhD is unavailable or it is unclear whether the scholar has completed a relevant PhD.

**Additional Data about Graduate Methods Syllabi**

In response to feedback that we received during the 2021 APSA meeting, we also collected data from a large corpus of methods syllabi associated with graduate-level political science programs. We collected the original syllabi from two main sources: (1) previous studies and data-collection efforts about graduate methods training (e.g., Emmons and Moravcsik 2019; Hardt et al. 2019); and (2) mass outreach to Directors of Graduate Studies in political science programs in the United States, Canada, and Europe. The syllabi in category (1) were principally associated with qualitative or multi-method courses; the syllabi in category (2) were principally quantitative.

For the latter effort, we initially contacted the Director of Graduate Studies at 103 political science PhD programs to request syllabi from quantitative methods courses taught between 1998 and 2018. A number of the respondents referred us to methods faculty in their department. In our follow up correspondence with the faculty, we requested referrals for additional contacts inside of the department. In total, we contacted 110 institutions and received quantitative syllabi from 62 institutions, with a response rate of 56 percent. We conducted this outreach from January – July 2022.

In total, our dataset captures methodological topics and pedagogical characteristics of 986 syllabi. After collecting and sorting the syllabi into a common archive, we collected the following categories of information from each syllabus. For all categories of variables indicated with an asterisk (\*), a value of 1 indicates that the syllabus includes information that aligns with the variable topic; a value of 0 indicates that the syllabus does not include this information; and a value of NA / NAV indicates that the syllabus does not have a discernible description of instructional topics. To justify a value of 1, we required that the syllabus devote at least part of one instructional week to the variable topic.

* Instructor details: We coded (1) the institution at which the instructor taught the course; (2) the instructor’s last name; (3) the course title; (4) the year and semester during which the instructor taught the course.
* General methodological tendency: We coded for the particular method of data collection or analysis associated with the course. We primarily used the course title and description to inform this category, even if some of the specific course material diverged from the method in question. We used the following categories, which correspond to our project’s general distinction between (1) quantitative, qualitative, and mixed / combined research approaches; and (2) methods of data collection and analysis:   
  + Quant General, combining both collection and analysis;
  + Quant Collection;
  + Quant Analysis;
  + Qual General;
  + Qual Collection;
  + Qual Analysis;
  + Mixed Collection, combining both Quant and Qual approaches;
  + Mixed Analysis; and
  + Unclear, including methods-inclined courses about the state of the political science discipline or the philosophy of social science.
* \* Pedagogy: We coded for whether the syllabus employed:
  + Exercises or problem sets, which we defined as specific, bounded, and detailed activities that practice key skills acquired in the course. In qualitative courses, practice exercises included: (1) interviews; (2) completing IRB applications; (3) completing Freedom of Information Act requests; (4) text analysis; (5) ethnography; and (6) tracing a process. We excluded activities that constitute a research design, including research-question specification, concept mapping, case-selection justification, literature reviews, and typologies. We also excluded grant proposals and prospecti.
  + Replication exercises, for which we required that the syllabus explicitly refer to the exercise as “replication.”
* \* Data collection: We coded for whether the syllabus included instruction about the following data-collection topics:  
  + Transformation of quantitative data, which we define as methods for changing or manipulating one or more pre-existing datasets collected by another party;
  + Collection of quantitative text data, which we define as methods for creating a numerical dataset from a pre-existing textual source;
  + Collection of quantitative interview data, which we define as methods for creating numerical data from interaction with other research participants, including observational surveys, survey experiments, and other experimental formats such as laboratory studies;
  + Collection of qualitative data, which we define as methods for creating non-numerical data from original sources such as archival research, interviews, focus groups, ethnography, participant observation, and oral histories;
  + Fieldwork, which we define as methods for collecting qualitative and quantitative data at a research site located away from a researcher’s home institution.
* \* Qualitative analytic methods: We coded for whether the syllabus included instruction about the following qualitative-methods topics, which we define in greater detail in our project codebook: (1) process tracing; (2) qualitative comparative analysis, or QCA; (3) congruence analysis; (4) counterfactual analysis; and (5) structured case comparison.
* \* Quantitative analytic methods: We coded for whether the syllabus included instruction about the following quantitative-methods topics, which we define in greater detail in our project codebook: (1) simple probability; (2) regression; (3) statistics with an identification strategy; and (4) machine learning.
* \* Research ethics: Although we do not analyze these data in this project, we also coded for whether the syllabus included instruction about the following topics in research ethics, or practices to minimize the potential individual and social harm of research activities, relationships, and institutions:  
  + Research design, including how researchers should determine the questions and approaches that they adopt in their research;
  + Data collection, including how researchers should interact with interlocutors and field sites;
  + Data analysis, including how and whether researchers should quantify complex political phenomena into numbers;
  + Positionality, including how and whether researchers should consider their social position---such as their gender, racial, or national identity---and its implications for their research process or findings;
  + Transparency, including how researchers should document their research process for broader accessibility and replicability;
  + Institutional Review Boards (IRB), including how researchers should evaluate and follow the institutional guidelines of IRBs;
  + Collaborative Institutional Training Initiative (CITI) certification, which refers to a requirement that students obtain CITI certification as a course requirement.
* Textbooks: Lastly, we coded author-year information about the specific textbooks on which the course relies. We only include references to book-length monographs that the course instructor describes at the top of the syllabus as a required or recommended instructional reference. We use NAV if the course syllabus does not indicate that the course uses a textbook.

**Data Analysis Procedures**

*Cleaning and Preparing the Data*

We exported the Media and Code Presence files from Dedoose into Excel. The Media file includes the title of each article uploaded, year and journal descriptors, as well as a variable for each code indicating the frequency with which it was applied to each article. The Code Presence file includes the title of each article and a binary variable for each code indicating application regardless of frequency. Both files included entries for the full sample of 1,926 articles. In Excel, the Media file was trimmed to include only those articles that were coded (n = 484).

We merged the two files in R. The resultant file included the article title, the year of publication, the journal title, and binary indicators for each code. We ran a series of checks on the file to ensure all articles were fully coded. After identifying all incomplete articles, the appropriate codes were applied in Dedoose. We then exported a new Code Presence file and merged it with the Media file. We repeated this process until all articles were deemed completely coded.

*Analyzing the Data*

We discuss each facet of our analysis – descriptive information about the articles in our sample, descriptive statistics about author and article parameters, and relationships among article parameters – in turn below. In each sub-section, we give a broad overview of how we analyzed the data in order to identify characteristics of the aggregate sample, and how we traced over-time change in particular parameters. For a more detailed description of the analysis please see our Pre-Analysis Plan.

# Articles in the Sample

*Number of articles published per year*: The count of all articles published in all journals included in the project in each year from 1998 to 2018.

*Number of articles published in each journal*: The count of all articles published in each journal in all years from 1998 to 2018.

*Number of articles published per year in each journal*: The count of all articles published in each journal, in each year from 1998 to 2018.

*Subfield*: We calculate the count and proportion of articles associated with (1) Comparative Politics (CP) exclusively; (2) American Politics (AP) exclusively; (3) International Relations (IR) exclusively; (4) Political Theory (PT) exclusively; (5) Methodology exclusively; (6) Conceptualization and Measurement exclusively; (6) CP and IR; (7) AP and IR; and (8) Methodology and all other subfield categories. To calculate the proportion of articles, we divide the count of all articles in each category by the total number of articles in the sample.

We also calculate the count and proportion of articles written by authors associated with each subfield in the APSA dataset. Because “self-identified subfield” and “article subfield” are different categories of subfield information, readers should not interpret this second calculation as a direct robustness check on the first.

*Number of articles in each subfield published in each journal*: The count of all articles published in each of the six subfields, in each journal.

# Descriptive statistics about author and article parameters

## A. Article-level attributes: Full sample

*Gender and authorship structure:* We calculate the count and proportion of articles that are (1) single-authored by a woman; (2) single-authored by a man; (3) co-authored by an all-women team; (4) co-authored by an all-men team; (5) co-authored by a team of women and men; (6) authored by any women, including single-authored, all-women, and mixed-gender teams; (7) authored by any men, including single-authored, all-men, and mixed-gender teams. To calculate the proportion of articles, we divide the count of all articles in each category by the total number of articles in the sample.

*Racial / ethnic identity and authorship structure*: We calculate the count and proportion of articles authored *only* by people who self-identify with the APSA racial / ethnic categories: (1) Non-Hispanic White or Euro-American; (2) Black, African American, or Afro-American; (3) East Asian or Asian American; (4) South Asian or Indian American; (5) Latino or Hispanic American; (6) Middle Eastern or Arab American; (7) Native American or Alaskan Native; (8) Native Hawaiian or Other Pacific Islander; plus an Other category that we create for all self-identified racial / ethnic categories that do not align with the above categories. These categories are not mutually exclusive because authors may either self-identify as multi-racial, or may self-identify as Hispanic or Latino (an ethnic category, according to the US census) and as members of other racial groups. We also calculate the count and proportion of articles authored by a combination of White and non-White people, to assess the extent of multi-racial / ethnic collaborations. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the sample for which we have information about the racial / ethnic identities of all authors. We omit all NA values from this calculation.  
  
*Gender by subfield, time, and journal*: We calculate the count and proportion of articles associated with the gender and authorship structure categories as they vary by (a) subfield; (b) time (year); and (c) journal. To calculate the proportion of articles, we divide the count of articles in each “crosstab” (e.g., the combination of gender and authorship structure and subfield) category by the total number of articles in the subfield, time, or journal categories for which we have information about the racial / ethnic identities of all authors.

## B. Author-level attributes: Full sample

*Gender self-identification*: Because we used author names to infer the author gender identities and authorship structure of each *article*, we also use data from the APSA member database as a measurement check on the gender identities of *authors* in the sample. First, we present the count and proportion of gender identities in the sample, differentiating between authors who self-identify (1) as men, (2) as women, and (3) as non-binary. Second, to facilitate this measurement check, we recreate the authorship structure measure from our original database. Based on the APSA database, we identify articles that have (1) only male authors; (2) only female authors; (3) only-male multi-author teams; (4) only-female multi-author teams; and (5) mixed-gender multi-author teams, including non-binary authors. We estimate the Pearson’s *r* on the relationship between the authorship structure measures based on our original inferences from author names, and the APSA self-identification data.

*Racial / ethnic identity*: Based on the APSA database, we present the count and proportion of authors who self-identify with the racial / ethnic categories that we describe above. To calculate the proportion of authors, we divide the count of authors in each category by the total number of authors in the sample for which we have information about the racial / ethnic identities of all authors. We omit all NA values from this calculation.

## C. Article-level attributes: Restricted sample

For the following attributes, we use a restricted sample of articles that excludes all articles that we coded as both (1) associated with the PT subfield and (2) having “No discernible method.” As we describe in Appendix A, these articles lack an empirical basis and a mode of analytic inquiry that corresponds to dominant methods in AP, CP, and IR. We exclude the PT / No discernible method articles from the restricted sample to avoid inflating the denominator (total articles in the sample) when calculating proportions of articles associated with different forms of data collection and analysis.

*Mode of inquiry*: We calculate the count and proportion of articles that drew on data the author generated (as opposed to exclusively using information collected from pre-existing sources) using (1) observational techniques; (2) experimental techniques; or (3) both observational and experimental techniques. Category (1) includes observational techniques involving both human participant and non-human participant-based forms of data collection. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.

*Empirical basis of analysis*: We calculate the count and proportion of articles that use data (1) only from pre-existing sources; (2) only from author-generated sources but excluding human participants; (3) only from author-generated sources, involving human participants with an observational sampling strategy; or (4) only from author-generated sources, involving human participants with an experimental sampling strategy. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.  
  
We also calculate the count and proportion of articles that use data from both pre-existing and author-generated sources. For these articles, we calculate the count and proportion that (1) tended towards relying on pre-existing sources or (2) tended towards relying on author-generated data. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the sample of articles that use data from both pre-existing and author-generated sources.

*Type of observational human*-*participant research*: We calculate the count and proportion of articles that use data from (1) ethnography or participant observation, exclusively; (2) ethnography or participant observation, in combination with any other data-collection technique; (3) interviews and focus groups, exclusively; (4) interviews and focus groups, in combination with any other data-collection technique; (5) survey methods, exclusively; or (6) survey methods, in combination with any other data-collection technique. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.  
  
We also calculate the count and proportion of articles associated with each category of observational human-participant research as they vary by (a) subfield; (b) time (year); and (c) journal. To calculate the proportion of articles, we divide the count of articles in each crosstab category by the total number of articles in the subfield, time, or journal categories.

*Type of experimental human*-*participant research*: We calculate the count and proportion of articles that use data from (1) survey experiments, exclusively; (2) survey experiments, in combination with any other data-collection technique; (3) field experiments, exclusively; (4) field experiments, in combination with any other data-collection technique; (5) lab experiments, exclusively; or (6) lab experiments, in combination with any other data-collection technique. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.  
  
We also calculate the count and proportion of articles associated with each category of experimental human-participant research as they vary by (a) subfield; (b) time (year); and (c) journal. To calculate the proportion of articles, we divide the count of articles in each crosstab category by the total number of articles in the subfield, time, or journal categories.

*Source of observational or experimental human-participant research*: We calculate the count and proportion of articles involving either observational or experimental human-participant research that drew on interaction with individuals (1) employed by international organizations or institutions; (2) employed by domestic governments; (3) employed by civil society organizations; (4) employed by media organizations; or (5) employed as academics or researchers. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.  
  
We also calculate the count and proportion of articles associated with each category of interlocutors as they vary by (a) subfield; (b) time (year); and (c) journal. To calculate the proportion of articles, we divide the count of articles in each crosstab category by the total number of articles in the subfield, time, or journal categories.

*Type of single-method analysis*: We calculate the count and proportion of articles that use only (1) interpretive methods; (2) qualitative methods, including (a) process tracing, (b) qualitative comparative analysis (QCA), (c) congruence analysis, (d) counterfactual analysis, (e) structured case comparison, or (f) another unspecified form of qualitative analysis; (3) illustrative case studies; (4) quantitative methods, including (a) simple probability, (b) regression, (c) statistics with an identification strategy, and (d) machine learning; (5) formal modeling; and (6) no discernible method. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.

*Type of multi-method analysis*: We calculate the count and proportion of articles that use (1) more than one method; (2) more than one qualitative method, excluding quantitative and interpretive methods, and formal modeling; (3) one or more qualitative methods and a quantitative method; (4) one or more qualitative methods and formal modeling; and (5) a quantitative method and formal modeling. To calculate the proportion of articles, we divide the count of articles in each category by the total number of articles in the restricted sample.  
  
For Categories (1), we calculate the count and proportion of articles that tended to (1) rely overwhelmingly on qualitative analysis; (2) rely overwhelmingly on formal modeling; (3) rely overwhelmingly on quantitative analysis; (4) rely overwhelmingly on interpretive methods; and (5) had no specific methodological focus. To calculate the proportion of articles, we divide the count of articles that tend towards each category of multi-method work by the total number of multi-method articles.  
  
For Categories (3) and (4), we also calculate the count and proportion of articles that tended to rely overwhelmingly on qualitative analysis. To calculate the proportion of articles, we divide the count of articles that tend towards each category of multi-method work by the total number of multi-method articles.

*Natural experiments*: We calculate the count and proportion of articles that claim to analyze data derived from a “natural experiment”. To calculate the proportion of articles, we divide the count of articles in the category by the total number of articles in the restricted sample. We also calculate the count and proportion of articles referencing a “natural experiment” as they vary by year. To calculate the proportion of articles, we divide the count of “natural experiment” articles published in each year by the total number of articles in the restricted sample published in the same year.

*Synthetic data*: We calculate the count and proportion of articles that use synthetic data as the empirical basis of the analysis. To calculate the proportion of articles, we divide the count of articles in the category by the total number of articles in the restricted sample. We also calculate the count and proportion of “synthetic data” articles as they vary by year. To calculate the proportion of articles, we divide the count of “synthetic data” articles published in each year by the total number of articles in the restricted sample published in the same year.

*Text mining and analysis*: We calculate the count and proportion of articles that use text mining or text analysis. To calculate the proportion of articles, we divide the count of articles in the category by the total number of articles in the restricted sample. We also calculate the count and proportion of “text mining / analysis” articles as they vary by year. To calculate the proportion of articles, we divide the count of “text mining / analysis” articles published in each year by the total number of articles in the restricted sample published in the same year.Analysis of relationships among article parameters

We also estimate a series of bivariate correlations and probit regressions to describe the relationships between multiple variables in the sample. We calculate Pearson’s *r* where we have no strong expectations about the direction of the “causal arrow” between two variables. Otherwise, we estimate both bivariate ordinary least squares (OLS) and probit regression models to characterize the relationship between more clearly defined independent and dependent variables. Because each outcome variable in our analysis is binary, the OLS and probit model specifications correspond to the potential linear or non-linear shape of the relationship, respectively. We present marginal effects for all probit models.

Consistent with convention, we use *α = 0.05* as our threshold for statistical significance for all models. We report p-values and unstandardized coefficients for all models.For most of the analyses, we restrict the analyses to exclude articles that we coded as associated with the PT subfield and having “No discernible method.” Where applicable we subset subsequent samples on the basis of this default sample.

1. We may engage in a separate subsequent analysis of the political theory scholarship in our sample through consultation with subfield scholars in our department. [↑](#footnote-ref-1)
2. When this occurred, the next article in the surplus that came from the same journal was used as a replacement. [↑](#footnote-ref-2)
3. Specifically, we consulted with Michael Bailey (Georgetown University), Andrew Bennett (Georgetown University), Markus Kreuzer (Villanova University), Alan Jacobs (University of British Columbia), Hans Noel (Georgetown University), and Jason Seawright (Northwestern University). We are grateful for their assistance. [↑](#footnote-ref-3)