How and Why Do Researchers Reference Data? A Study of Rhetorical Features and Functions of Data References in Academic Articles

RESEARCH PAPER

SARA LAFIA
ANDREA THOMER
ELIZABETH MOSS
DAVID BLECKLEY
LIBBY HEMPHILL

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data reuse is a common practice in the social sciences. While published data play an essential role in the production of social science research, they are not consistently cited, which makes it difficult to assess their full scholarly impact and give credit to the original data producers. Furthermore, it can be challenging to understand researchers' motivations for referencing data. Like references to academic literature, data references perform various rhetorical functions, such as paying homage, signaling disagreement, or drawing comparisons. This paper studies how and why researchers reference social science data in their academic writing. We develop a typology to model relationships between the entities that anchor data references, along with their *features* (access, actions, locations, styles, types) and *functions* (critique, describe, illustrate, interact, legitimize). We illustrate the use of the typology by coding multidisciplinary research articles (n = 30) referencing social science data archived at the Inter-university Consortium for Political and Social Research (ICPSR). We show how our typology captures researchers' interactions with data and purposes for referencing data. Our typology provides a systematic way to document and analyze researchers' narratives about data use, extending our ability to give credit to data that support research.

CORRESPONDING AUTHOR:

Sara Lafia

ICPSR, University of Michigan, Ann Arbor, MI, USA

slafia@umich.edu

KEYWORDS:

citation analysis; data reuse; research assessment; scholarly communication

TO CITE THIS ARTICLE:

Lafia, S, Thomer, A, Moss, E, Bleckley, D and Hemphill, L. 2023. *How and Why Do Researchers Reference Data?* A Study of Rhetorical Features and Functions of Data References in Academic Articles. *Data Science Journal*, 22: 10, pp. 1–15. DOI: https://doi.org/10.5334/dsj-2023-010

1. INTRODUCTION

As datasets enter the scientific record, citations connect published data to a larger research network (Hey et al. 2009). Data citations establish precedence for results, provide evidence signaling the quality and significance of research findings, and make it possible to study how researchers use existing data. Citation analysis relies upon the standardization of citations to assess scholarly communication patterns, such as the reach or visibility of ideas across scientific disciplines (e.g., through paper citation networks) (Lafia et al., 2022). Citation indexes of academic papers, like the Science Citation Index (Garfield 1964), allow researchers to understand who is highly cited, which published work is highly cited, and which publication outlets are prominent. Recent audits of bibliometric networks have also revealed inequalities, suggesting that citations are not objective (Kwon 2022). For instance, citation rates vary by research topic and author race and gender, suggesting that social factors play an important role in researchers' awareness of published work and decisions to cite it (Kozlowski et al. 2022).

While the infrastructure for studying citation trends for research publications is robust, three main challenges limit the large-scale analysis of data citations. The first challenge is the unambiguous identification of data references. While there are well-established systems for referencing the work of others (Chernin 1988), many authors still fail to cite data. Many authors refer to data informally in their writing, despite data repositories' guidance on best practices for data citation (Fenner et al. 2019) and pressure from funders and publishers to 'make data count' (Cousijn et al. 2019). When the burden of linking data to publications falls largely on the author, this often results in partial or inconsistent references to datasets in research articles (Boland et al. 2012). Informal data citation practices make it challenging for readers to understand which data the authors accessed and whether they analyzed data or simply described them (Moss and Lyle 2018).

A second challenge involves understanding the intent of a data citation. Bibliometric analysis often treats citations as something that can be standardized and universally interpreted as conferring legitimacy to published work (Cronin 1981). However, like citations of academic literature, researchers cite data for different purposes. Existing citation typologies account for the variety of reasons that researchers cite materials (e.g., to persuade, to critique, to contrast). Many researchers communicate their findings through empirical studies in which they make claims tied to other scholarly products, including published data. Authors' claims range in specificity from explicit to implicit and are often supported by data (Blake 2010).

A third related challenge involves inferring the quality of citations. Bibliometric measures for quantitative impact assessment, such as the h-index, indicate the popularity or visibility of a source (Egghe 2010) but say little about the nature of engagement surrounding it. Prior studies of citations to academic literature distinguish surface citations from those that engage deeply with the source material (Cronin 1984; Leydesdorff 1998; Spiegel-Rosing 1977; White and Wang 1997) and help determine the purpose or polarity of citations (Abu-Jbara et al. 2013; Cohan et al. 2019; Hernández-Alvarez and Gomez 2016; Teufel et al. 2006). For example, citations that pay homage (e.g., to one's mentors or other influential researchers) also create cumulative advantages, where the best-known researchers receive far more credit for their work (Merton 1968). Thus, the number of citations a source has received does not indicate the purpose of the citations and may not be a reliable proxy for research quality (Garfield 1979).

Given the challenges associated with analyzing data references, this study takes a qualitative approach to identify the types of, reasons for, and interactions involving social science data reuse in scientific research. Prior studies of data reference have focused on formal bibliographic citation (Belter 2014; Jiao and Darch 2020; Mooney and Newton 2012; Park et al. 2018). By contrast, we closely analyze even oblique data *mentions* in papers—sentences in which a dataset, or part of a dataset, is named but not formally cited. We find that data references perform a limited set of functions, which we define in a typology that captures and describes the variety of ways researchers refer to data. We then apply the typology to analyze the use of research data in social science publications.

2. BACKGROUND

2.1 DEFINING AND CITING DATA

The terms 'data' and 'datasets' have various meanings depending on their context. Often, what becomes 'data' is determined by scientists' choices as they interact with and record observations. 'Data,' then, are a byproduct of interpretation and can be practically understood as 'referring only to that which is analyzed' (Coombs 1964). Part of this challenge in defining 'data' relates to their 'unruly' and 'poorly bounded' identities, which makes it difficult for them to function as digital objects that can be readily referenced and retrieved (Wynholds 2011). There is also disagreement on the use of the term 'dataset' in technical and scientific literature, which presents challenges for data sharing and preservation (Renear et al. 2010). 'Data' are abstract arrangements of symbols that express content; 'datasets' are made up of multiple data-bearing entities and may contain additional contextual information about data, such as collection methods (Furner 2016; Wickett et al. 2012). In our analysis, we focus on archived social science datasets that include contextual metadata and documentation, which have been produced and shared by others for research purposes.

Capturing the relationships between datasets and other scholarly works is critical for giving credit to datasets. Data 'references,' mentions,' and 'citations' signal importance and enable credit through attribution (Altman et al., 2015). While the terms 'references' and 'citations' are often used interchangeably, 'references' generally indicate that the work is listed in the reference section of a publication (Gilbert and Woolgar 1974). The term 'citation' implies the use of a persistent identifier (PID), which carries a more formal connotation in bibliometrics than 'references' or 'mentions' (Ball and Duke 2015). Citations with PIDs link published works to their usage contexts, enabling the verification and reuse of existing scientific analyses (Buneman et al. 2022; Shotton 2010). While authors are beginning to use digital object identifiers (DOIs) to reference data, best practices for when and why they should do so are not widely followed (Mayo et al. 2016). For example, a recent review of literature citing datasets using their DOIs found that many authors cited data that they had mentioned or described (e.g., data collection methods) but which had not been re-analyzed or used in other ways (Banaeefar et al. 2022). In our work, we use the general term 'data reference' to cover informal data mentions (i.e., use of a dataset name only), formal citations (i.e., the inclusion of an APA-style citation for a dataset DOI), and descriptions of data use.

2.2 CITATIONS IN SCHOLARLY COMMUNICATION

Citation analysis within scientific disciplines reveals information flows and brings together separate strands of information to construct 'consensus models' of subjects within science (Garvey and Griffith 1972). Citations reflect influences on authors, and citation patterns trace communication across active research networks (Edge 1979). Citations function differently at the micro and macro levels. At the micro level, citations indicate professional relations and function as rewards, while at the macro level, groups of citations function as concept symbols that codify knowledge in hierarchical social networks (Leydesdorff 1998). When cited, papers can be invoked as symbolic of the ideas expressed in their text (Small 1978). Citations are powerful in that they are persistent and take on a separate identity from the people involved in their creation. They are 'speech acts,' which are brief statements that endure in documents and can be inspected over time (Smith 2014). Cited sources often substantiate statements or assumptions, point to further information, acknowledge previous research in the same area, and draw critical comparisons indicating the quality of the research (Spiegel-Rosing 1977).

Studies of research infrastructure rely on citations as metrics for tracing attribution and indicating the impact of scholarly works like datasets or software (Mayernik et al. 2017). Institutions, like journals and data publishers, enforce disciplinary and cultural norms for writing style and citation through publishing guidelines and style manuals. Data citations that use specific identifiers allow readers to identify, retrieve, and give credit to research data. Despite recommendations and best practices, formal data citation is still not commonplace in scholarly writing (Mooney and Newton 2012). Incomplete, informal, or improperly formatted citations present obstacles to tracking data use (Zhao et al. 2018). Vague or implicit references, for example, make it difficult for readers without an intimate knowledge of variables or other data features to understand which data the authors used and how they used them (Moss and

Lyle 2018). Thus, focusing exclusively on formal citation practices (e.g., using DOIs) means overlooking many potential data references. We seek a more comprehensive understanding of what authors do rhetorically when they formally and informally refer to data in their papers.

2.3 MEANING AND MOTIVATIONS FOR CITATION

Citations bestow credit and recognition in science (Cronin 1984). However, there may be a disconnect between authors' citation practices and the use of citations to evaluate performance and measure research impact. In other words, citations indicate what is cited and how often but do not explain 'why' works are cited. Authors' motivations for citing publications can be classified as scientific or tactical. Scientific citations provide background, identify gaps, and establish bases for comparison. Tactical citations acknowledge subjective norms and advertise published work (Lyu et al. 2021). While it is often assumed that citations indicate high-quality work that has influenced authors' research, a survey found that authors' citation decisions were more often motivated by strategic factors rather than their familiarity with the research or perceptions of quality (Teplitskiy et al. 2022).

Behavioral surveys and interviews with authors reveal their judgments about what they are citing and why, which are not reflected in the scholarly record (Liu 1993). In one such study, authors considered the recency of publications, their topical specificity, and ease of use when deciding whether to cite them (White and Wang 1997). Authors also believe that citations reflect the prominence or novelty of a document as a 'concept marker' and that citing the concept marker will bolster the authority of one's work, either through alignment or by critiquing existing work (Case and Higgins 2000). Silvello identified six main motivations for data citations that are shared across scientific fields: data attribution (accountability and merit), data connection (to claims in publications), data discovery (identification and retrieval), data sharing (reputational), data impact (assessing exposure), and reproducibility (validation and procedures) (Silvello 2018). These motivations alone, however, do not explain authors' data-citing behaviors and why they vary across venues and contexts.

2.4 ANALYZING CITATION CONTENT AND CONTEXT

Many computational approaches for citation analysis have been proposed, building on prior insights about authors' motivations to cite. Common citation categories identified across multiple content analysis studies included background information, theoretical framework, prior empirical or experimental evidence, negative distinction, and explanation of methodology (Ding et al. 2014). Features, such as the sections of publications in which citations appear, can be used along with the semantic content of citations to predict citation intent (Nakov et al. 2004). Various classification schemes have been proposed for labeling authors' intents in citing published research (Hernández-Alvarez and Gomez 2016). One such scheme accounts for citation purposes (i.e., author intent) and polarity (i.e., author sentiment) by distinguishing and weighting negative, neutral, and positive citations (Abu-Jbara et al. 2013). More granular, rule-based coding schemes differentiate statements of weakness, contrasts, or comparisons with other work, agreement, compatibility with other work, and neutral citations (Teufel et al. 2006). Conversely, less granular schemes support general citation intent classification by distinguishing background information, method, and comparison citations (Cohan et al. 2019). Such schemes can also help distinguish citation framing (e.g., uses, motivation, future, extends, compare or contrast, background) (Jurgens et al. 2018).

Qualitative approaches respond to the difficulty of predicting authors' intents by focusing on the context and features of references. A review of data citations in academic literature found that they varied along two major dimensions: cited entities and styles (Fear 2013). Data producers (the researchers who created the data) and data providers (the people or the institution from which the data were obtained) were often named in data citations. Another recent study found that researchers tended to use data created by others for comparison (e.g., ground-truthing, calibration, and identifying baseline measures) and integration (e.g., to ask new questions and conduct new studies) (Pasquetto et al. 2019). A large survey found that existing data are often used as the basis for a new study, to prepare for a new project, to generate new ideas, to develop new methods, to verify other data through analysis and sensemaking, and for teaching (Gregory et al. 2020). Given that data reuse fulfills diverse needs, researchers' purposes for

citing data also vary. We build upon established insights into data reuse practices to show how researchers give credit to data.

3. MATERIALS AND METHODS

3.1 SAMPLING FRAME

We analyzed a sample of publications referencing one or more datasets available through the Inter-university Consortium for Political and Social Research (ICPSR), a large social science data archive at the University of Michigan. We based our typology on existing citation schemes for academic publications, which we extended and refined through iterative coding. The typology captures structural features and rhetorical functions that authors employ when referencing research data. Prior studies have focused on particular publication styles, such as data papers (Jiao and Darch 2020; Li and Jiao 2022), or publication outlets, such as PLoS One (Zhao et al. 2018). Instead, we drew from multi-disciplinary publications that referenced social science data archived at ICPSR. This approach allowed us to capture a wider variety of data reference contexts. We considered data references as they occurred in the full-length context of research publications. Further, our only selection requirement was that each publication mentioned one or more archived social science datasets.

We analyzed papers retrieved as part of ICPSR's collection efforts to expand the ICPSR Bibliography of Data-related Literature. The Bibliography includes more than 100,000 publications that use existing social science data available through ICPSR. The review process for the Bibliography involves searching bibliographic databases for references to data available through published ICPSR studies. Staff manually review the metadata and full text of publication search results for evidence of data use. ICPSR maintains strict collection criteria to ensure that publications in the Bibliography reflect data use. Publications are collected if they unambiguously refer to one or more studies available through ICPSR and if it is clear that the authors have accessed and analyzed the data. Publications are rejected from the Bibliography if they fail to demonstrate substantial use of ICPSR data or if the specific studies or series used in the authors' analyses cannot be determined.

To develop a sampling frame for testing our typology, we first identified five publications from the current ICPSR Bibliography representing the multidisciplinary use of ICPSR data. We closely read these publications to identify data references and develop a provisional typology. We then searched an external index of publication full text provided by the Dimensions bibliometric database (Hook et al. 2018) for additional references to any of ICPSR's 11,639 study DOIs available as of February 2022. With the support of ICPSR Bibliography staff, we evaluated and classified the 2,546 search results into six categories indicating whether the publications met the collection criteria for the ICPSR Bibliography. These categories were proposed by ICPSR staff (Banaeefar et al. 2022) and are defined in **Appendix A (Supplementary File 1)**. We then randomly selected publications across each category to include in our analysis, resulting in a total of thirty publications. We gathered additional metadata for each publication, such as the field of research categories from Dimensions, to determine the disciplinary coverage of our sample. We report the publication sampling frame and selection criteria in **Appendix A (Supplementary File 1)**.

3.2 QUALITATIVE CODING

Our team conducted two phases of coding. The purpose of the first phase was to develop a codebook to describe the diversity of data references. To develop and refine our codes, three annotators from our team read the full-text multiple times for each publication labeled 'phase I' listed in **Appendix A (Supplementary File 1)**. The annotators independently proposed refinements to a shared version of the codebook. To achieve qualitative reliability, the annotators drew from many examples and discussed them in weekly meetings (Bauer 2000). The annotators met weekly to review and incorporate proposed changes in interactive coding sessions. Emerging ideas and conflicting opinions created a dialog from which the codebook was created. In addition to codes, annotators also discussed the scope of the data references and the codebook's focus, purpose, and definitions. Each team member independently applied the updated codes for every iteration of the codebook to identify and annotate all data references

in the full text. The team repeated this process until saturation was reached and no new codes were proposed (Charmaz 2006). The first coding phase resulted in a stable codebook, which we report in **Appendix B (Supplementary File 2)**.

We also described the extent to which annotators' coding aligned. Given that the annotators selected and coded segments from the unstructured, full text of publications, we used Holsti's Index (1969) as an agreement measure. The final Holsti Index was 67.9%, indicating a relatively high level of agreement, given that annotators selected different text segments, which they coded with multiple codes. A single team member independently applied the typology to the held out set of twenty-five publications labeled 'phase II' in the sampling frame reported in **Appendix A (Supplementary File 1)**. This second phase demonstrated the typology in action and captured findings shared in **Section 4**.

4. RESULTS

The data reference typology consists of four parent codes (*Data Entity*, *Data Reference*, *Feature*, and *Function*), which are summarized and defined in Table 1. A *Data Entity* anchors a *Data Reference* and is based on the pragmatic distinctions raised in work by Renear et al. (2010) to define components of datasets in scientific literature. Renear et al. distinguish between data content (including files and observations), groupings (the set or study to which they belong), and purposes (metadata used to interpret the data). Similarly, we define *Data Entities* as 'one or more words indicating recorded observations' and record them as 'Files,' 'Metadata,' 'Studies,' or 'Variables.' We excluded *Data Entities* that were not specific or were not discussed in the body of the paper. For example, if the name of a dataset appeared in the title of a publication but data were not described in the main text, we did not consider this a *Data Entity*. Given that some data analysis discussions were broad (e.g., results of statistical tests), we only considered statements that referenced a specific *Data Entity*.

Table 1 Overview of parent codes, subcodes, and definitions.

PARENT CODE	SUBCODES		DEFINITION
	1ST LEVEL	2ND LEVEL	
Data Entity	File		One or more words indicating recorded observations
	Metadata		
	Study		
	Variable		
Data Reference			Context window in which one or more *Data Entities* are mentioned
Feature	Access	Provision, Reception	Structure, form, and appearance of the data reference
	Action	Cites, Mentions, Uses	
	Location	Abstract, Acknowledgements, Appendix, Caption…	
	Style	Acronym, Generic, Name, Parenthetical	
	Type	Derived, Primary, Secondary	
Function	Critique	Comparison, Limitations	The purposes of the data reference
	Describe	Composition, Source	
	Illustrate	Context, Outlook	
	Interact	Interpretation, Manipulation	
	Legitimize	Justification, Transparency	

A *Data Reference* is the context window in which one or more *Data Entities* appear. We experimented with various context windows and determined that paragraphs captured sufficient detail leading up to and following a *Data Entity*. We focused on three types of data references, which are introduced in **Section 2.1**: *citations*, *mentions*, and *uses*. *Data citations*

include a clear pointer to a published data source but do not name a dataset or indicate that authors used the data. *Data mentions* name a dataset in order to describe it but do not indicate data use. *Data uses* name a dataset and describe interactions between the author and the data. We applied feature and function codes to each *Data Reference*. The full codebook, along with definitions, rules, and examples are provided in **Appendix B (Supplementary File 2)**.

4.1 FEATURES OF DATA REFERENCES

Features describe the structure, form, and appearance of the *Data Reference*. The twenty-four features of data references that we identified are organized under five subcodes (*Access*, *Action*, *Location*, *Style*, and *Type*). *Access* codes indicate whether the author describes data sharing or retrieval in the reference. *Action* codes capture the distance between the author and the data along a continuum covering 'citing' (i.e., parenthetically referencing data without further context), 'mentioning' (i.e., describing or alluding to data), and 'using' (i.e., describing active, hands-on work with data). The *Action* codes build on the distinction proposed by Pasquetto et al. (2019) between comparative and integrative data reuse. The *Location* code notes the section of the publication in which the *Data Reference* occurs, such as the abstract, acknowledgments, captions, figures, tables, footnotes, or methods sections stated in the expanded IMRAD structure (Sollaci and Pereira 2004). The *Style* code captures how the author specifies data entities through the use of an acronym such as 'ANES,' a generic noun such as 'data' or 'study,' a formal name such as the 'American National Election Study, 2016,' or a parenthetical citation using author and year of publication. The *Type* code captures whether the data entity is derived from existing data, represents a primary source created by the authors, or is a secondary source published for other researchers to use.

4.2 FUNCTIONS OF DATA REFERENCES

Functions reflect the purposes of each *Data Reference*. The ten rhetorical functions of data references we identified are organized into five subcodes (*Critique*, *Describe*, *Illustrate*, *Interact*, and *Legitimize*). Definitions and examples for each code and subcode are provided in **Appendix B (Supplementary File 2)**.

The *Critique* code includes (1) *Comparison*, which contrasts the author's work with other work that uses the data or findings from other sources. Authors issue *Comparisons* to draw a contrast between their work and prior findings or to summarize conclusions drawn from the prior use of data. The *Critique* code also includes (2) *Limitations*, which signal authors' awareness and caution when working with data. This code includes acknowledging quality issues, such as potential errors or sampling biases that should limit how data are used.

The *Describe* code includes (3) *Composition*, which explains or discusses knowledge about the data or metadata. *Composition* data references describe what is in the data (e.g., the sampled population) or the study's context (e.g., the data collection method). The *Describe* code also includes (4) *Source*, in which authors describe the provenance of the data. *Source* references acknowledge the origin of the data associated with the data producer or provider.

References labeled with the *Illustrate* code are persuasive. This code includes (5) *Context*, in which the author provides background, findings, or statistics derived from referenced data. In a *Context* reference, the data are metonymic, standing in for the point or claim that the authors are making. The *Illustrate* code also includes (6) *Outlook*, in which authors speculate on potential applications of data that they did not conduct or review in their work. *Outlook* references claim the potential utility of data based on their properties.

The *Interact* code describes hands-on work with data. *Interact* includes a subcode for (7) *Interpretation*, where authors make an empirical claim derived from the analysis of referenced data. *Interpretation* references often follow the description of the authors' analysis. The *Interact* code also includes a subcode for (8) *Manipulation*, where authors describe steps performed while working with data. *Manipulation* involves selecting variables, preparing or transforming data for analysis, and specific data preparation techniques like sampling, correlating, integrating, and validating analyses.

Finally, the *Legitimize* code is used for references intended to persuade the reader through value statements made about data. The *Legitimize* code includes (9) *Justification*, which draws

attention to a feature of data that lends credibility or authority to the authors' choices. Examples of *Justification* references include authors' reasons for why data were selected, discussions about the credibility or representativeness of data, and descriptions of previous data uses that qualify their selection. The *Legitimize* code also includes a subcode for (10) *Transparency*, in which authors explain why or how an analysis procedure was applied and signal quality or considerations taken in the analysis. Examples of *Transparency* references include making methods open or reproducible by including analysis in supplementary materials.

4.2.1 Data references provide readers with access to data

Data references provide multiple ways of accessing data. Though some journals require that authors include data *provision* statements, where authors make the data used in their analyses available to readers, they were not common in the publications we reviewed. Examples that we encountered included cases where authors provided access to data derived from their analysis for the stated purpose of replication. Alternatively, authors may also provide access to data as a means of recruiting future collaborations. One such description of data *provision* read:

> The authors have made available the data that underlie the analyses presented in this article (see Styck, Beaujean, & Watkins 2019), thus allowing replication and potential extensions of this work by qualified researchers. Next users are obligated to involve the data originators in their publication plans, if the originators so desire (Styck et al. 2019).

Statements about authors' access, or *reception*, of data from providers were often accompanied by formal, parenthetical data references. We defined data *reception* as a reference to an existing data entity and specifications for how that data could be accessed. For example, if a reference is parenthetical, the instance in the reference list must provide an access mechanism, such as a URL, by which others may access the source. In the following example, the author formally attributes the data creator through a parenthetical citation, which includes details about the analysis performed and the historical context motivating selection of the dataset:

> In order to test whether or not fallout from nuclear testing had persistent effects on the agricultural sector, I create a panel of comparable variables from Historical U.S. Agricultural Censuses for the years 1940 to 1997 Haines et al. (2015). This Census data comes from the most comprehensive surveys of agriculture in the United States that ranges back to 1840. Starting in 1920, the Agricultural Census started conducting bidecennial surveys. I use this data to explore the effects on radioactive fallout deposition on long run outcomes and agricultural development at a national level (Meyers 2019).

4.2.2 Data references indicate authors' interactions with data

Data references spanned three levels of interactions between authors and data. First, we identified examples of superficial data *citations*, where authors' cited published datasets in the same way as academic articles. In these cases, authors did not name a specific dataset in their writing; instead, they used footnotes or parenthetical citations to formally acknowledge the dataset in their reference list. Most data citations were found in introductory sections and were contextual, meant to provide background, findings, or statistics, which authors used to substantiate a point. It was often unclear, however, how statistics or figures that the authors cited were connected to or derived from the source data. In the following example, the data citation provides findings without direct analysis. No verbs have been used to describe actions performed with or to data; instead, the reader may assume that the authors have some previous experience analyzing the data or that the cited figure is tied to the dataset's published summary statistics. In the following example, the author provides statistics with a corresponding footnote, which leads to a formal citation for data from the India Human Development Survey in the article's reference list:

> Slums are associated with poor quality housing, water, sanitation, and other services, leading to, among other outcomes, higher rates of disease and death. Rich households, on the other hand, are often located in areas with piped water and during water shortages can build storage facilities, tap into underground wells, and

> pay for delivered water. Only 38% of households among the poorest fifth of India's urban population have access to indoor piped water compared with 62% of the richest fifth (Frumkin et al. 2020).

When authors *mentioned* data, instead of *citing* them, they described the composition or source of a dataset. Unlike citations, *mentions* name the data in-line. We identified mentions of data primarily in the articles' Methods, Introduction, Discussion, and footnotes. Many data *mentions* provided details about the composition of the data product and relayed knowledge about the basis for the study, collection method, or population. Mentioning data provided background information about data that the authors used later in their analysis or acknowledged the authors' awareness of data that they evaluated but decided not to use. In the following example, the authors describe changes made to the sampled population between waves of a survey in order to qualify their selection method, signal awareness of data quality, and justify their approach:

> As regards education, health, relationship status, and employment status, Wave 1 respondents who did not remain within the analytical sample show disadvantages compared with those who did. Accordingly, if those more susceptible to depressive symptoms had lower likelihoods of remaining within the analytical sample, attrition between Waves 1 and 2 might lead to conservative assessments of how contexts undergoing economic declines affect their residents' depressive symptoms (Settels 2021).

4.2.3 Data references are building blocks for empirical arguments

In examples where data were *critiqued*, authors described others' prior efforts or findings to contrast with their approaches. In some cases, authors described how they used the same data differently or decided against using the data based on the reasons that they provided. An example of a data comparison is provided below. The authors present several longitudinal studies covering a similar population and explain potential differences in findings based on differences in their compositions. In this way, the authors signal that they have performed due diligence; they are aware of related studies and can describe their limitations:

> Studying an earlier cohort than After the JD, the National Longitudinal Bar Passage Study found that long-term bar passage rates were substantially lower for minorities than for whites. Thus a study of all law degree holders including those who did not pass a bar examination may find larger racial gaps in earnings. Census surveys such as those used in this paper lack bar passage status, and therefore likely include a larger proportion of lower earning individuals compared to After the JD (McIntyre and Simkovic 2018).

More references indicating data *use* were found in Methods and Discussion sections of articles as well as in captions, figures, and tables. Mentions and data use statements were distinguished based on the authors' use of verbs and personal pronouns. Most of the use statements described actions, specifically data manipulation (e.g., steps performed while working with data) and interpretation (e.g., making an empirical claim derived from data analysis). Data references describing *use* also occurred in appendixes and supplementary materials rather than in designated areas of articles, like acknowledgments or data availability statements. Examples of data manipulation included selecting variables from referenced data and preparing, transforming, modifying, sampling, subsetting, comparing, or correlating referenced data. Data interpretation included building theories, comparing, and interpreting empirical evidence in figures. The following example illustrates how an author refers to two waves of a study, and related variables, in detailing their analytical approach:

> To assess the degree to which genetic and environmental factors are stable over time requires an extension of the classical twin design to encompass repeated measurements. Here, we used the bivariate Cholesky decomposition approach: for each of n measured variables, the Cholesky decomposition specifies n latent A, C, and E factors. Viewed as a diagram, with the latent factors arranged above the measured variables, each of these factors is connected to the measured (manifest) variable beneath it, and to all variables to the right. In this way, each latent factor is

connected to one fewer variables than the preceding factor. This design is of value for answering the current question as it allows estimation both of A, C, and E effects at Wave 1, and the extent to which these can account for Wave 2 variance, as well the new variance that emerges at Wave 2 (Lewis and Bates 2017).

5. DISCUSSION

Our typology expands the notion of data use beyond re-analysis. For example, while some researchers may access and re-analyze published survey data, many more may reuse that survey's questionnaire or sampling design as a gold standard. Users may also critique the survey data by pointing to its limitations in addressing a particular topic. Our approach casts a wide net to capture these kinds of data references, providing insights into how social science data support research. Our typology is useful for informing recommendation scenarios for researchers about when, why, and how they should reference published data. It also provides a basis for novel data reuse metrics that reflect many forms of engagement with data, from the reuse of survey designs to the re-analysis of survey data.

Our typology also reveals some ways in which data references differ from and align with traditional bibliographic citations. First, the referenced entity can vary in scale; we found references to individual files, metadata records, studies overall, and individual variables. While bibliographic citations may similarly range in scale (e.g., a citation of a specific phrase or section of a paper vs. the paper overall), data entities have a different and possibly broader range of constituent parts. Further expansion and refinement of the typology through review of papers in other domains may reveal additional sub-entities (for instance, research in archaeology or paleontology likely refer to specific artifacts, as well as data derived from those artifacts). Further work is needed to understand the implications of these differing citation scales; are different scales (e.g., variable-level versus full dataset-level) references associated with different types of use and argumentation? Are different scales of data more or less likely to result in a formal citation of the dataset? Data entities may additionally have multiple versions that could be referenced (though we did not see this in our sample); how does this complicate our ability to trace the flow of scholarly influence?

Second, we find that data references can act as 'concept symbols' (Leydesdorff 1998), similarly to bibliographic references. Informal reference to datasets by acronym or name (and without a formal citation) indicates a familiarity with datasets as one sees with canonical works of scholarship. In other words, datasets can be referenced with the same familiarity as a biologist references Darwin or an economist references Locke. Future work to identify these foundational or canonical datasets may help reveal how datasets-as-concept symbols differ from bibliographic references. Datasets may be unique in that they also can have a distinct metonymic function, where a reference to a dataset as a whole can stand in for a reference to a specific part or feature of a dataset (as revealed by our 'context' code).

Third, data references show interactions with data entities that aren't typically found with bibliographic entities—namely, the provision and archiving of data. Datasets function as both a resource to be used, and a scholarly product to be cited or made available to others. In the publications we annotated, we found that it was uncommon for authors to provide direct access to the findings that they derived from existing data; more often, authors established credibility and trust by simply describing the data source or data provider that they had accessed. Prior studies of researchers' attitudes toward data sharing and reuse show that researchers are reluctant to provide access to their data because they do not believe that the data would be valuable to others, or because hoarding data provides a way to attract future collaborations (Cragin et al. 2010; Pasquetto et al. 2019). Though our sample was not representative, we found early indications that align with this prior work.

Finally, we also found alignment with prior schemes describing authors' motivations for citing literature. The rhetorical functions we identified signal the quality, verifiability, or reproducibility of authors' research findings by allowing readers to discover the data the authors have analyzed (Silvello 2018). For the most part, the data references we reviewed either provided details about dataset composition or descriptions of data manipulation. In the examples we identified, authors affixed additional context about the analysis they performed to connect a

data source to its use. Further, when authors included specific access information for data, this enabled readers to retrieve the same dataset.

5.1 LIMITATIONS AND FUTURE WORK

This study proposes a typology that models how authors reference research data. We developed the typology by closely reading papers from the ICPSR bibliography and adding new categories until we reached saturation. The present analysis is not intended to provide quantitative evidence for specific citation trends. We would need to conduct annotation at a larger scale with additional measures in place to verify the agreement of annotators. In addition, we constructed our sampling frame by selecting papers that were first reviewed and classified by experts (i.e., ICPSR Bibliography staff); the sample is balanced across the categories provided in **Appendix A (Supplementary File 1)**. While this sampling strategy is useful for developing and analyzing the ICPSR Bibliography, future uses of the typology for other purposes may require different selection criteria.

We envision applying our typology to study differences in data references across social science disciplines (e.g., sequences or co-occurrences of data reference strategies as markers of scientific disciplines or analytical methods). A recent study of data citation practices at ICPSR observed unexpected uses of dataset DOIs in published literature, which did not indicate data use (Banaeefar et al. 2022). Our typology can be used to study when and why researchers use dataset DOIs and distinguish references that describe data from those that imply data analysis.

6. CONCLUSION

Although research data are increasingly important in modern scientific analyses, they have not been regarded historically as primary research products. The publication, long-term preservation, and dissemination of research data, along with descriptive metadata, make it possible for others to discover, use, and cite observations collected by other researchers for other purposes. We introduced a typology of data references that characterizes the functions data serve in scientific publications: critical, descriptive, illustrative, interactive, and legitimizing. The typology captures researchers' interactions with (e.g., work or analyses done with data) and judgments about data (e.g., claims about its fitness for use based on what is known about data). Understanding why authors reference research data is essential for giving data producers and providers the scholarly research credit they deserve for facilitating scientific work.

DATE ACCESSIBILITY STATEMENT

The ICPSR Bibliography of Data-related Literature is available at https://www.icpsr.umich.edu/web/pages/ICPSR/citations/.

ADDITIONAL FILES

The additional files for this article can be found as follows:

- **Appendix A (Supplementary File 1).** Sampling Frame. DOI: https://doi.org/10.5334/dsj-2023-010.s1
- **Appendix B (Supplementary File 2).** Codebook. DOI: https://doi.org/10.5334/dsj-2023-010.s2

ACKNOWLEDGEMENTS

We thank Elizabeth Yakel, Morgan Wofford, Lizhou Fan, and Bethany Radcliff from the University of Michigan School of Information for their comments on earlier drafts.

FUNDING INFORMATION

This material is based upon work supported by the National Science Foundation under grant 1930645. This project was made possible in part by the Institute of Museum and Library Services LG-37-19-0134-19.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualization, A.T., S.L., E.M., D.B., and L.H.; Formal Analysis, S.L., D.B., and E.M.; Funding Acquisition, L.H., and A.T.; Methodology, S.L., E.M., D.B., A.T., and L.H.; Writing - Original Draft, S.L., L.H, A.T., D.B., and E.M.

AUTHOR AFFILIATIONS

Sara Lafia orcid.org/0000-0002-5896-7295
ICPSR, University of Michigan, Ann Arbor, MI, USA

Andrea Thomer orcid.org/0000-0001-6238-3498
School of Information, University of Arizona, Tucson, Arizona, USA

Elizabeth Moss orcid.org/0000-0001-5464-8716
ICPSR, University of Michigan, Ann Arbor, MI, USA

David Bleckley orcid.org/0000-0001-7715-4348
ICPSR, University of Michigan, Ann Arbor, MI, USA

Libby Hemphill orcid.org/0000-0002-3793-7281
ICPSR, University of Michigan, Ann Arbor, MI, USA; School of Information, University of Michigan, Ann Arbor, MI, USA

REFERENCES

Abu-Jbara, A, Ezra, J and **Radev, D.** 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, June 2013, 596–606. Association for Computational Linguistics. Available at: https://aclanthology.org/N13-1067.

Altman, M, Borgman, C, Crosas, M and **Matone, M.** 2015. An introduction to the joint principles for data citation. *Bulletin of the American Society for Information Science/ASIS*, 41(3): 43–45. Wiley Online Library. DOI: https://doi.org/10.1002/bult.2015.1720410313

Ball, A and **Duke, M.** 2015. How to cite datasets and link to publications. DOI: https://doi.org/10.1007/1-4020-5340-1

Banaeefar, H, Burchart, S, Moss, E, et al. 2022. Best practice may not be enough: Variation in data citation using DOIs. *Poster presented at the annual meeting of the International Association for Social Science Information Service and Technology*, June 9, 2022. DOI: https://doi.org/10.7302/4809

Bauer, MW. 2000. Classical content analysis: A review. In: *Qualitative Researching with Text, Image and Sound: A Practical Handbook*. Sage, 131–151. DOI: https://doi.org/10.4135/9781849209731

Belter, CW. 2014. Measuring the value of research data: a citation analysis of oceanographic data sets. *PloS one*, 9(3): e92590. DOI: https://doi.org/10.1371/journal.pone.0092590

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

Boland, K, Ritze, D, Eckert, K, et al. 2012. Identifying references to datasets in publications. In: *Theory and Practice of Digital Libraries*, 2012: 150–161. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-33290-6_17

Buneman, P, Dosso, D, Lissandrini, M, et al. 2022. Data citation and the citation graph. *Quantitative Science Studies*, 2(4): 1399–1422. MIT Press – Journals. DOI: https://doi.org/10.1162/qss_a_00166

Case, DO and **Higgins, GM.** 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*. Wiley Online Library. DOI: https://doi.org/10.1002/(SICI)1097-4571(2000)51:7<635::AID-ASI6>3.0.CO;2-H

Charmaz, K. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Chernin, E. 1988. The 'Harvard system': a mystery dispelled. *BMJ: British Medical Journal*, 297(6655). BMJ Publishing Group: 1062. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1834803/ (accessed 15 February 2023). DOI: https://doi.org/10.1136/bmj.297.6655.1062

Cohan, A, Ammar, W, van Zuylen, M, et al. 2019. Structural scaffolds for citation intent classification in scientific publications. In: *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N19-1361

Coombs, CH. 1964. A theory of data. Oxford, England: Wiley A theory of data. Available at: https://psycnet.apa.org/fulltext/1965-00053-000.pdf.

Cousijn, H, Feeney, P, Lowenberg, D, et al. 2019. Bringing citations and usage metrics together to make data count. *Data science journal,* 18. Ubiquity Press, Ltd. DOI: https://doi.org/10.5334/dsj-2019-009

Cragin, MH, Palmer, CL, Carlson, JR, et al. 2010. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926): 4023–4038. The Royal Society Publishing. DOI: https://doi.org/10.1098/rsta.2010.0165

Cronin, B. 1981. The need for a theory of citing. *The Journal of documentation; devoted to the recording, organization and dissemination of specialized knowledge*, 37(1): 16–24. Emerald. DOI: https://doi.org/10.1108/eb026703

Cronin, B. 1984. *The citation process: The role and significance of citations in scientific communication*. Taylor Graham.

Ding, Y, Zhang, G, Chambers, T, et al. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9): 1820–1833. Wiley. DOI: https://doi.org/10.1002/asi.23256

Edge, D. 1979. Quantitative measures of communication in science: a critical review. *History of science; an annual review of literature, research and teaching,* 17(36 Pt 2): 102–134. DOI: https://doi.org/10.1177/007327537901700202

Egghe, L. 2010. The Hirsch index and related impact measures. *Annu. Rev. Inf. Sci. Technol.,* 44(1): 65–114. DOI: https://doi.org/10.1002/aris.2010.1440440109

Fear, KM. 2013. *Measuring and anticipating the impact of data reuse*. Available at: https://deepblue.lib.umich.edu/handle/2027.42/102481.

Fenner, M, Crosas, M, Grethe, JS, et al. 2019. A data citation roadmap for scholarly data repositories. *Scientific data,* 6(1): 28. DOI: https://doi.org/10.1038/s41597-019-0031-8

Frumkin, H, Das, MB, Negev, M, et al. 2020. Protecting health in dry cities: considerations for policy makers. *BMJ*, 371: m2936. DOI: https://doi.org/10.1136/bmj.m2936

Furner, J. 2016. 'Data': The data. In: Kelly, M and Bielby, J (eds.), *Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro*. Wiesbaden: Springer Fachmedien Wiesbaden, 287–306. DOI: https://doi.org/10.1007/978-3-658-14681-8_17

Garfield, E. 1964. 'Science Citation Index'—A New Dimension in Indexing. *Science,* 144(3619): 649–654. American Association for the Advancement of Science (AAAS). DOI: https://doi.org/10.1126/science.144.3619.649

Garfield, E. 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics,* 1(4): 359–375. Springer Science and Business Media LLC. DOI: https://doi.org/10.1007/BF02019306

Garvey, WD and **Griffith, BC.** 1972. Communication and information processing within scientific disciplines: Empirical findings for Psychology. *Information Storage and Retrieval*, 8(3): 123–136. DOI: https://doi.org/10.1016/0020-0271(72)90041-1

Gilbert, GN and **Woolgar, S.** 1974. Essay review: The quantitative study of science: An examination of the literature. *Science studies,* 4(3): 279–294. SAGE Publications. DOI: https://doi.org/10.1177/030631277400400305

Gregory, K, Groth, P, Scharnhorst, A, et al. 2020. Lost or found? Discovering data needed for research. *Harvard Data Science Review*. DOI: https://doi.org/10.1162/99608f92.e38165eb

Haines, M, Fishback, P and **Rhode, PW.** 2015. United States Agriculture Data, 1840–2010. Dataset ICPSR35206-v2, Inter-university Consortium for Political and Social Research. Ann Arbor, MI.

Hernández-Alvarez, M and **Gomez, JM.** 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering,* 22(3): 327–349. DOI: https://doi.org/10.1017/S1351324915000388

Hey, T, Tansley, S and **Tolle, K.** 2009. *The fourth paradigm: Data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

Holsti, OR. 1969. *Content analysis for the social sciences and humanities*. Reading. MA: Addison-Wesley.

Hook, DW, Porter, SJ and **Herzog, C.** 2018. Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3. DOI: https://doi.org/10.3389/frma.2018.00023

Jiao, C and **Darch, PT.** 2020. The role of the data paper in scholarly communication. *Proceedings of the Association for Information Science and Technology,* 57(1). Wiley. DOI: https://doi.org/10.1002/pra2.316

Jurgens, D, Kumar, S, Hoover, R, et al. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6. MIT Press: 391–406. DOI: https://doi.org/10.1162/tacl_a_00028

Kozlowski, D, Larivière, V, Sugimoto, CR, et al. 2022. Intersectional inequalities in science. *Proceedings of the National Academy of Sciences of the United States of America*, 119(2). DOI: https://doi.org/10.1073/pnas.2113067119

Kwon, D. 2022. The rise of citational justice: how scholars are making references fairer. *Nature*, 603(7902): 568–571. DOI: https://doi.org/10.1038/d41586-022-00793-1

Lafia, S, Fan, L, Thomer, A and **Hemphill, L.** 2022. Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. *Quantitative Science Studies*, 3(3): 694–714. DOI: https://doi.org/10.1162/qss_a_00209

Lewis, GJ and **Bates, TC.** 2017. The temporal stability of in-group favoritism Is mostly attributable to genetic factors. *Social Psychological and Personality Science*, 8(8): 897–903. DOI: https://doi.org/10.1177/1948550617699250

Leydesdorff, L. 1998. Theories of citation? *Scientometrics,* 43(1): 5–25. Springer Science and Business Media LLC. DOI: https://doi.org/10.1007/BF02458391

Li, K and **Jiao, C.** 2022. The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *Journal of the Association for Information Science and Technology,* 73(6): 834–846. Wiley. DOI: https://doi.org/10.1002/asi.24585

Liu, M. 1993. Progress in documentation the complexities of citation practice: a review of citation studies. *Journal of Documentation,* 49(4): 370–408. MCB UP Ltd. DOI: https://doi.org/10.1108/eb026920

Lyu, D, Ruan, X, Xie, J, et al. (2021) The classification of citing motivations: a meta-synthesis. *Scientometrics,* 126(4): 3243–3264. DOI: https://doi.org/10.1007/s11192-021-03908-z

Mayernik, MS, Hart, DL, Maull, KE, et al. 2017. Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology,* 68(6): 1341–1359. Wiley. DOI: https://doi.org/10.1002/asi.23721

Mayo, C, Vision, TJ and **Hull, EA.** 2016. The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository, 11(1): 150–155. DOI: https://doi.org/10.2218/ijdc.v11i1.400

McIntyre, F and **Simkovic, M.** 2018. Are law degrees as valuable to minorities? *International Review of Law and Economics*, 53: 23–37. DOI: https://doi.org/10.1016/j.irle.2017.09.004

Merton, RK. 1968. The Matthew effect in science. The reward and communication systems of science are considered. *Science*, 159(3810): 56–63. Available at: https://www.ncbi.nlm.nih.gov/pubmed/5634379. DOI: https://doi.org/10.1126/science.159.3810.56

Meyers, K. 2019. In the shadow of the mushroom cloud: Nuclear testing, radioactive fallout, and damage to U.S. agriculture, 1945 to 1970. *The journal of economic history,* 79(1): 244–274. Cambridge University Press. DOI: https://doi.org/10.1017/S002205071800075X

Mooney, H and **Newton, MP.** 2012. The anatomy of a data citation: Discovery, reuse, and credit. *Journal of librarianship and information science*. DOI: https://doi.org/10.7710/2162-3309.1035

Moss, E and **Lyle, J.** 2018. Opaque data citation: Actual citation practice and its implication for tracking data use. Available at: https://deepblue.lib.umich.edu/handle/2027.42/142393.

Nakov, PI, Schwartz, AS, Hearst, M, et al. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In: *Proceedings of the SIGIR*, 2004: 81–88. Citeseer. DOI: https://doi.org/10.1525/rep.2004.88.1.81

Park, H, You, S and **Wolfram, D.** 2018. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, 69(11): 1346–1354. Wiley. DOI: https://doi.org/10.1002/asi.24049

Pasquetto, IV, Borgman, CL and **Wofford, MF.** 2019. Uses and reuses of scientific data: The data creators' advantage. *1.2* 1(2). MIT Press – Journals. DOI: https://doi.org/10.1162/99608f92.fc14bf2d

Renear, AH, Sacchi, S and **Wickett, KM.** 2010. Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1): 1–4. Wiley. DOI: https://doi.org/10.1002/meet.14504701240

Settels, J. 2021. Compound disadvantage between economic declines at the city and neighborhood levels for older Americans' depressive symptoms. *City & community*, 20(3): 260–288. SAGE Publications. DOI: https://doi.org/10.1177/1535684120980992

Shotton, D. 2010. CiTO, the Citation Typing Ontology. *Journal of biomedical semantics*, 1(Suppl 1): S6. DOI: https://doi.org/10.1186/2041-1480-1-S1-S6

Silvello, G. 2018. Theory and practice of data citation. *Journal of the Association for Information Science*. Wiley Online Library. DOI: https://doi.org/10.1002/asi.23917

Small, HG. 1978. Cited documents as concept symbols. *Social studies of science*, 8(3): 327–340. SAGE Publications Ltd. DOI: https://doi.org/10.1177/030631277800800305

Smith, B. 2014. Document acts. *Institutions, Emotions, and Group Agents*, 19–31. DOI: https://doi.org/10.1007/978-94-007-6934-2_2

Sollaci, LB and **Pereira, MG.** 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association: JMLA*, 92(3): 364–367. Available at: https://www.ncbi.nlm.nih.gov/pubmed/15243643.

Spiegel-Rosing, I. 1977. Science studies: Bibliometric and content analysis. *Social studies of science*, 7(1): 97–113. SAGE Publications Ltd. DOI: https://doi.org/10.1177/030631277700700111

Styck, KM, Beaujean, AA and **Watkins, MW.** 2019. Profile reliability of cognitive ability subscores in a referred sample. *Archives of Scientific Psychology,* 7(1): 119–128. DOI: https://doi.org/10.1037/arc0000064

Teplitskiy, M, Duede, E, Menietti, M, et al. 2022. How status of research papers affects the way they are read and cited. *Research policy*, 51(4): 104484. DOI: https://doi.org/10.1016/j.respol.2022.104484

Teufel, S, Siddharthan, A and **Tidhar, D.** 2006. Automatic classification of citation function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, July 2006, 103–110. Association for Computational Linguistics. Available at: https://aclanthology.org/W06-1613. DOI: https://doi.org/10.3115/1610075.1610091

White, MD and **Wang, P.** 1997. A qualitative study of citing behavior: Contributions, criteria, and metalevel documentation concerns. *The Library quarterly*, The University of Chicago Press. 67(2): 122–154. DOI: https://doi.org/10.1086/629929

Wickett, KM, Sacchi, S, Dubin, D, et al. 2012. Identifying content and levels of representation in scientific data. *Proceedings of the American Society for Information Science and Technology*, 49(1): 1–10. Wiley. DOI: https://doi.org/10.1002/meet.14504901199

Wynholds, L 2011. Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International journal of digital curation*, 6(1): 214–225. Edinburgh University Library. DOI: https://doi.org/10.2218/ijdc.v6i1.183

Zhao, M, Yan, E and **Li, K.** 2018. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 69(1): 32–46. Wiley. DOI: https://doi.org/10.1002/asi.23919

TO CITE THIS ARTICLE:
Lafia, S, Thomer, A, Moss, E, Bleckley, D and Hemphill, L. 2023. *How and Why Do Researchers Reference Data?* A Study of Rhetorical Features and Functions of Data References in Academic Articles. *Data Science Journal*, 22: 10, pp. 1–15. DOI: https://doi.org/10.5334/dsj-2023-010

Submitted: 16 February 2023
Accepted: 03 April 2023
Published: 28 April 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.