Check for updates

Journal of Management
Vol. 49 No. 3, March 2023 875–880
DOI: 10.1177/01492063221141861
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions

Improving Our Field Through Code and Data Sharing

Timothy J. Quigley
University of Georgia
Aaron D. Hill
University of Florida
Andrew Blake
Texas Tech University
Oleg Petrenko
University of Arkansas

Across academia, there is a series of ongoing and intertwined debates about the need for research transparency, greater confidence in the accuracy of empirical findings, and the overall relevance and credibility of our work (for a summary, see Bergh, Sharp, Aguinis, & Li, 2017). Central to these debates is the idea that, as a field, we need greater confidence in the cumulative body of scientific knowledge created by our research. To make progress toward this goal, we need processes that minimize the publication of flawed results emanating from honest errors; insufficient training; and, in the worst of cases, fraud.

Many stakeholders are pressing for initiatives related to these challenges. For example, some journals, and those responsible for their oversight, debate the merits of code and data sharing, whereas others are implementing policies that encourage or even require such practices (for a review, see Dosch & Martindale, 2020). Similarly, some grant-awarding organizations require a level of code and/or data transparency in exchange for funding (for a discussion and examples, see Global Innovation Fund, 2021; Metzenbaum, 2021). There is also a growing recognition that replications are integral to the knowledge generation process (e.g., Köhler & Cortina, 2021). Advocacy organizations have also arisen to facilitate the warehousing of code and data related to research efforts (e.g., FIVES Project; Center for Open Science).

In this editorial, the *Journal of Management* has asked us to discuss and describe a related but slightly different approach. As a means of making the research process more efficient,

Corresponding author: Timothy J. Quigley, University of Georgia, C210 Benson Hall, Athens, GA 30602, USA.

E-mail: tquigley@uga.edu

findings more robust and reliable, and the results and implications of our studies more trustworthy, we propose that academics cultivate ongoing collaborations to create shared code and data libraries that can benefit the field. Doing so will provide numerous benefits to researchers while also making better use of the tax dollars, philanthropic contributions, and student tuition that ultimately fund our research endeavors.

Collaborative Approach

The management field has a distinguishing focus that values the efficient use of resources. Researchers working in organizational behavior, for example, seek to understand how the interactions of human beings in work groups can be configured to enhance efficiency. In strategic management, there is a desire to understand how firms can efficiently configure organizational resources to maximize various economic and stakeholder-oriented outcomes. Yet, as academics studying topics related to the efficient use of resources, we often fail to put those insights into practice in our own work. To be specific, common external databases such as Compustat, Execucomp, and CRSP (Center for Research in Security Prices) are often combined and processed to generate base samples and variables used in many of our studies. Similarly, datasets from different survey waves (each typically having multiple questions tapping into various constructs) are often merged and processed with many computations to arrive at the sample and measures used in the final analyses. Extensive code, written in various software packages or languages (e.g., R, Stata, Python), capture the needed logic to merge and process data from raw files to final analysis. The resulting code can often stretch to thousands of lines.

Take, for example, the process of assembling a core dataset of variables relevant to studies of firm-level outcomes in publicly traded firms. Most studies in this domain create an initial sample from firms covered by Compustat (which captures financial information from annual reports). If the study involves aspects of the top management team or executive compensation, data from BoardEx or Execucomp must be accessed, processed, and merged with the base dataset. If the study requires the inclusion of securities analyst ratings, then individual ratings from the Institutional Brokers Estimate System (IBES) must be aggregated and merged into the core dataset. The need for board-level data (Risk Metrics or BoardEx), mergers and acquisitions data (SDC Platinum), or corporate social performance data (KLD) might necessitate half a dozen or more similar steps to complete the dataset. Each step requires not only complicated code to process and merge data, but also manual checks to address missing data and fix underlying data inconsistencies. Scholars must also make judgment calls to address common anomalies in the data. Accordingly, each step, from accessing data to the final reporting of results, provides many opportunities to introduce error or bias into the research process.

Despite the complexity of these common data merging and processing tasks, many scholars at this moment are performing duplicative work to accomplish the same outcome. Sure, data merging and processing can, at times, be done with a few dozen or maybe a few hundred lines of code. At the same time, if you are reading this thinking, "that's all it took me the last time I did it," then your resulting dataset likely has flaws of unknown consequence on your results. Knowing this, one might ask why such duplication of effort happens. One might also ponder how many of these efforts include undetected errors that yield imprecise

measures or incomplete data merges that result in biased samples and flawed inferences. Ultimately, we must consider how this collection of both within-study mistakes and between-study inconsistencies affects the body of empirical work produced by our field.

To address such problems and put into practice our field's research on the efficient use of resources, we advocate for a shift in the mindset of management scholars toward a philosophy where code and data are routinely shared, evaluated, and improved upon by the community. We envision such sharing could come in at least three different forms that arise from processes common to many data collection efforts, as follows:

1. ***Data assembly code sharing***. Code to perform routine tasks such as processing, merging, and preparing publicly available data for analysis (or processing survey responses for common measures) could be converted into generic modules and made public and available for improvement by the academic community. An example of this would be the code needed to link the CEO and their annual compensation from Execucomp to annual firm-level data in Compustat.
2. ***Data correction routine sharing***. Routines or repositories developed to correct data errors or account for strange anomalies in large datasets could also be made public and available for improvement by the academic community. For example, building on the case above, Execucomp often fails to identify a CEO for a firm, identifies more than one person as CEO without details on who served when, or has missing or inaccurate information about start and end dates. In some cases, the CEO of a subsidiary is mistakenly noted as the CEO of the entire firm. We can envision shared code and data patches that resolve this type of inconsistency.
3. ***Unique variable sharing***. Data that might be useful for other studies, perhaps even as a control that might be otherwise difficult to obtain, such as forced CEO turnover (Gentry, Harrison, Quigley, & Boivie, 2021), could be made public and available for enhancement through collaboration.

In short, any code and data collection or structuration process that might be common across studies could be shared and improved rather than kept proprietary. Doing so would reduce errors, increase transparency, and decrease the time needed to complete aspects of the research process. Notably, these processes could happen independently from the sharing of final code and data that may be required by some journals following the publication of an article. That is, we are calling for sharing code and data related to creating datasets rather than the code and data needed to reproduce final results.

A reasonable argument against our proposed approach might be the time and effort needed to prepare code and data libraries for public release. Here, we believe there are two intertwined benefits of our approach that will result in a net time savings. First, having existing code libraries will reduce the overall time it takes to create datasets for future studies. In turn, scholars using them will only need to document additions or changes to existing code. Second, by working from standard code libraries, students training in the field will have better examples from which to work and, by extension, will be better trained in how to generate code that will make their own efforts more efficient and, in time, also benefit the field.

There are tools in place that are designed to facilitate the type of sharing we advocate. GitHub, for example, is a large repository that allows for the open storage of and collaboration on code and data and is already used by many academics. RunMyCode.org is a website that "enables scientists to openly share the code and data that underlie their research publications." Further, repositories such as the Duke University Research Data Repository, FIVES Project, and the recently started Circles of the Strategic Management Society can serve as warehouses for code and datasets. Together, these and other existing tools like them provide readymade platforms for scholars to engage in the transparent and collaborative behavior we suggest.

Example and Benefits

As noted in a special announcement in the March 2022 issue of the *Journal of Management,* the authors of this editorial recently launched the SMART tool (Standardized Measures that are Accurate, Replicable, and Time-saving), which is available at www.smartdatatool.net. The SMART tool aims to resolve issues common to processing and merging archival data in management research (such as that found in Compustat, Execucomp, and CRSP), thus offering a resource that both exemplifies the ideals we espouse in this editorial and enables researchers to streamline a four-part database preparation process of:

1. ***Data download***. Standardized code that downloads and links Wharton Research Data Services (WRDS) datasets.
2. ***Measurement catalog***. A centralized catalog of popular measures in the field of strategic management, including definitions and mathematical formulas for each.
3. ***Patches***. Fixes for common errors and missing data in the underlying datasets.
4. ***Structuration code***. Standardized code that can be executed using open-source software R to generate a dataset that is reproducible and storable for reuse, modification, or extension for specific applications.

Specifically, the SMART tool serves as an initial step toward the kind of transparent and collaborative sharing of code and data we call for here. That is, using the SMART tool provides a standardized means of merging commonly used datasets; allows for the inclusion of both raw data from each of these datasets and a range of calculated measures that frequently appear in the management literature; and provides a "data patch" that fixes common problems in the RiskMetrics board member database. As an example of how the SMART tool and others like it could expand, a forthcoming data patch will provide corrections to the CEO start and end dates in Execucomp. Others will find issues not addressed by existing data patches and can submit them for inclusion. In time, additional datasets, manipulations, measures, and data patches can be incorporated through collaborations among varied members of the research community.

We envision a number of benefits from the SMART tool and others like it. First, a common code base allows for standardized measurement. To illustrate, consider that Hopkins and Lazonick (2016) highlighted eight different measures of total compensation in Execucomp and demonstrated how associated relationships and conclusions drawn from these can vary drastically depending on which measure is used. The SMART tool, and others like it, can help add accuracy and facilitate replication by providing standardized measurement and

labeling across studies. Relatedly, such tools can streamline the generation of multiple measures to assess the robustness or boundary conditions of findings. Further, the quick generation of samples and common measures can allow reviewers to easily check that descriptive statistics presented in a manuscript match expectations.

Second, collaborative code and data sharing tools improve measurement accuracy by helping avoid unintentional errors and even intentional manipulation in measurement calculations. That is, by using calculations that are displayed publicly and validated for accuracy, the SMART tool and those like it avoid errors tied to human fallibility (e.g., a typographical error in a calculation). At the same time, the public nature of the code aids in comparing means and standard deviations of data reported in a study by simplifying examinations for inaccuracies (which should occur during the review process). Providing multiple, accurate measures also facilitates comparisons across studies and offers an avenue to more accurate conclusions, both within and between studies.

Third, even with the best efforts of authors and reviewers to establish clarity, certain measures can be difficult to explain, and human error may occur. Collaboration efforts such as the SMART tool afford a means for improved clarity and replicability by pointing to a specific measure with consistent labeling, transparent mathematical formulas, and the specific code used in computation.

Fourth, on a practical level, the process of sourcing and matching data and subsequently calculating measures can be time intensive and, while necessary, not particularly value-added for science (especially in cases when the process is frequently duplicated). Scholars might be better suited to focus their valuable time and limited resource in other ways. The SMART tool and other related collaborative efforts free up time, allowing scholars to pursue opportunities more central to our collective purpose. In addition, reviewers, knowing that collaborative resources are readily available and will not be overly time restrictive to use, may be more willing to ask for additional tests using standardized data and measures that can enhance the robustness and reliability of our research as well as to cross-check accuracy for themselves.

Finally, amid debates about the relevance of academic research and questions about allocating taxpayer, philanthropic, and student tuition funds for such purposes, tangible efforts to collaborate on code and data allow scholars to generate reliable scientific findings more quickly. Further, time previously allocated to repetitive tasks can be refocused toward more valued activities and may help demonstrate responsiveness to these important stakeholder concerns.

Conclusion

As technology advances to ease collaboration, we call for a commensurate shift in the mindset of management scholars toward a philosophy where the code and data used to create, aggregate, and clean datasets are routinely shared, evaluated, and improved upon by the community. The shift we call for can happen separately from, and without the potential problems associated with, any journal-specific requirements to share code and data following publication. We believe such a shift is both feasible, given the current technology, and necessary as our academic community moves toward a new paradigm where the expectations for transparency, reliability, and sharing become more widespread.

Senior scholars, who enjoy the benefits of tenure, should lead the way in these efforts by sharing their own code and data while infusing a different mindset in the next generation of

scholars currently being trained in PhD programs across the field. Done right, responding to our call might even hasten the progression toward sharing of code and data associated with papers by reducing some of the related concerns of that practice. In turn, the benefits of code and data sharing may help academics show we are being responsive to important questions about the reliability of our research.

ORCID iDs

Timothy J. Quigley https://orcid.org/0000-0002-1077-569X
Aaron D. Hill https://orcid.org/0000-0002-9737-1718

References

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. 2017. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization*, 15: 423-436.

Dosch, B., & Martindale, T. 2020. Reading the fine print: A review and analysis of business journals' data sharing policies. *Journal of Business & Finance Librarianship*, 25: 261-280.

Gentry, R. J., Harrison, J. S., Quigley, T. J., & Boivie, S. 2021. A database of CEO turnover and dismissal in S&P 1500 firms, 2000–2018. *Strategic Management Journal*, 42: 968-991.

Global Innovation Fund 2021. Guidelines on research transparency and ethics in GIF-related impact evaluation. https://www.globalinnovation.fund/wp-content/uploads/2021/12/Research-Transparency-and-Ethics-Guidelines-2021.pdf.

Hopkins, M., & Lazonick, W. 2016. The mismeasure of mammon: Uses and abuses of executive pay data. *Institute for New Economic Thinking Working Paper Series,* 49. https://www.ineteconomics.org/research/research-papers/the-mismeasure-of-mammon-uses-and-abuses-of-executive-pay-data.

Köhler, T., & Cortina, J. M. 2021. Play it again, Sam! An analysis of constructive replication in the organizational sciences. *Journal of Management*, 47: 488-518. doi:10.1177/0149206319843985

Metzenbaum, S. H. 2021. *Federal grants management: improving transparency* [White paper]. IBM Center for The Business of Government. https://www.businessofgovernment.org/sites/default/files/Improving%20Transparency.pdf.