SYMPOSIUM

Data Collection, Opportunity Costs, and Problem Solving: Lessons from Field Research on Teachers' Unions in Latin America

Christopher Chambers-Ju, *University of California, Berkeley*

As I prepared to leave Mexico, my efforts to access an important dataset for my dissertation became increasingly desperate. A newspaper article mentioned that the Mexican teachers' union endorsed nearly 2,000 political candidates in the 2012 election.[1] I wanted to solicit this list. For three weeks I made daily treks to the Mexican teachers' union's headquarters in downtown Mexico City. The union was in a dark, cavernous building. I wandered through smoked-filled offices with impressionistic portraits of the union's president, Elba Esther Gordillo. I sought "Geraldo," a leader of the union's political action committee, who could tell me whether the data I sought was obtainable, or give me closure with a simple "no."[2]

One day I received news that Geraldo could meet. After hurrying to his office, the receptionist told me that her boss had just stepped out to take an urgent call. Confused, I sat and waited. When it became clear that Geraldo would not appear, I left. Later I was told what had happened. Apparently, after agreeing to the interview, Geraldo informed the union's top brass that I wanted to know about the union's political endorsements. Because of a recent damaging leak to the press, Geraldo was ordered not to speak to me, and I was warned to stop contacting union leaders. A gatekeeper had slammed the door on my research.

This article examines the challenges related to getting data in the field. Researchers can be confused about what to do when data exists but, because of bureaucratic barriers, cannot be accessed; researchers may face challenges just "getting into the building" (Jensenius this symposium). Rich data may be found, but putting it together may require significant work—the amount of effort may outweigh the data's added value. Comparative researchers may encounter data that is asymmetric and only available for some cases, and they must decide whether to present incomplete data in their final analysis. This article lays out these challenges and then suggests paths forward.

The approach to field research outlined in this article is informed by my experience in Argentina, Colombia, and Mexico studying the electoral participation of teachers' unions. I wanted to analyze how teachers came together as an organized voting bloc, whether teachers influenced the vote choice of low information voters, why union leaders became political candidates, and what conditions enabled teacher-based parties to form. I collected electoral data from official government sources, teacher surveys, newspaper archives, and databases of legislator CVs. I draw from my experience throughout this article to illustrate general points with specific examples.

THE COST STRUCTURE OF DATA COLLECTION

Field research promises high rewards because original data can make valuable contributions to the field.[3] Graduate students may postpone graduation because of these payoffs. By conducting interviews and communicating directly with the actors involved in generating data, analysis is sharpened and made more credible. There is broad agreement that more data is better than less and that more types of data—qualitative and quantitative—are better than fewer. However, data collection involves labor, capital, and time. Empirically minded social scientists almost always encounter problems related to data collection. Analysts tend to specialize in collecting and analyzing only a few types of data—be they interview transcripts, survey data, or others—because of the costliness of working with various data structures. Many activities in the field have a high marginal cost because of the costs associated with conducting searches, which ex ante are unknowable.[4] For instance, the per-unit cost of conducting an additional face-to-face interview involves contacting, scheduling, preparing questions, transportation, carrying out the interview, and then transcription, coding, and analysis. The marginal cost of one more interview is quite high.

Field research is one of the most costly forms of data collection, especially at the beginning of a project and early in an academic career. Graduate students face opportunity costs; they could instead learn new quantitative methods or graduate early. Data collected for theory building is open ended. Few guidelines are available about what is needed and what is not. Usually, a significant chunk of the data that is collected is not used in the final analysis.[5] Graduate students are expected to soak and poke, and then quickly move to a research question and an empirical strategy. They must commit to a project before knowing exactly what it is or whether it will work.

Gathering data in a foreign country also adds a dimension of complexity. There may be different levels of access depending on country context; developing countries present considerable hurdles. Not all countries have freedom of information acts and laws requiring transparency and disclosure. Working with data in a foreign language is also more challenging.

doi:10.1017/S1049096514000304

Gathering data in multiple foreign countries compounds these problems. As the complexity of a project increases, something is likely to go wrong.

DATA COLLECTION PROBLEMS

The costliness of data collection gives rise to various problems. Projects motivated by important questions often run into a gap between the research design and the available data. Ideally, analysts will find data that is structured. Yet, analysts often encounter data that is nonexistent, unstructured, hidden, or asymmetric. Table 1 presents a typology of data that analysts may encounter in the field, with a brief description of the challenges that each poses.

The costliness of data collection gives rise to various problems. Projects motivated by important questions often run into a gap between the research design and the available data. Ideally, analysts will find data that is structured. Yet, analysts often encounter data that is nonexistent, unstructured, hidden, or asymmetric.

Structured Data

Structured data serves as a standard against which more complicated types of data can be compared. Structured data is complete, clean, and ready to be analyzed. Interuniversity Consortium for Political and Social Research has complete datasets that are publicly available. Similarly, government agencies also can be transparent and have easily downloadable data. For my research, I found that electoral data was easily accessible through the websites of government agencies. Getting data on the territorial distribution of the vote share of candidates and parties linked to the teachers' union proved relatively straightforward.

Nonexistent Data

Important research questions can call for data that does not exist. Nonexistent data is a problem commonly encountered when conducting historical research. It is more problematic if the analyst has reason to believe that data, in fact, exists, and significant time must be spent to verify that it does not. When there is evidence to suggest that data is nonexistent, the analyst must make decisions; whether to search for alternative indicators, or proxies, that can indirectly get at the research question, or whether to generate new data, for instance, through a survey. Generating original data can be orders of magnitude more costly than collecting data.

For my research, I sought the results of elections for the Colombian teachers' union's executive committee—for 1994 and 1998—to understand how union electoral results influenced the decisions of union leaders to pursue public office. Several helpful union leaders told me that the union lacked "institutional memory;" union leaders had not archived this data and it did not exist. I chose to move on and seek out other data that would be more useful.

Unstructured Data

Data—in some form—exists for most significant social and political phenomena that have occurred over the past 20 years; most leave an electronic footprint. The problem, however, is that data is scattered among various sources and is quite messy. The more disorganized data is, the higher the costs of putting it together. For example, when data is found in newspaper articles, it must be collected from various sources, coded, and then aggregated. Data for different years or for different regions may be housed in various locations. For my research, putting together a census of union leaders who went into public office required me to organize unstructured data. First, I searched databases with legislator CVs to find former union leaders who entered the national legislature. This data, however, was incomplete. I then printed tables with the names of union leaders who had gone into public office and circulated these tables in interviews. I asked current union leaders to verify that the legislators I had found did, in fact, have ties to the union and to identity other union leaders who had gone into government as well as friends of the union. I would then check these names online. This iterative approach yielded interesting data but constructing this dataset required me to invest a significant amount of time consulting multiple sources.

Hidden Data

Data known to exist may be hidden. Finding and accessing it can be difficult, and a lot of time can be wasted in fruitless

Table 1

Types of Data Problems

TYPE OF DATA	DESCRIPTION
Structured	Data exists, is clean, and is ready to be analyzed
Non-Existent	Data does not exist. Alternative indicators must be found or new data must be generated.
Hidden	Data exists, but it is difficult to access.
Unstructured	Data exists, but it is costly to collect and clean.
Asymmetric	Data exists, but it is partial or incomplete for key cases; it is difficult to analyze.

https://doi.org/10.1017/S1049096514000304 Published online by Cambridge University Press

searches. Some of this data may be accessible through networking, creativity, and luck; others may be nearly impossible to access, short of paying large sums of money. The candidate endorsement data I sought from the Mexican teachers' union is a good example of hidden data. After being told to stop contacting union leaders, I decided to move on and search for low-hanging fruit.

Asymmetric Data

Analysts conducting comparative research frequently encounter asymmetric or incomplete data. Data may be available and complete for one or more cases, but nonexistent for others. Data can also be coded or operationalized in different ways, rendering it non-comparable. Asymmetric data raises difficult questions that must be addressed in the final analysis. In some cases, asymmetric data is sufficiently interesting that it can be presented in an incomplete form.[6] Yet, data with too many holes must be thrown out. Comparative researchers must make decisions about how much asymmetry their analysis can tolerate.

Teacher surveys were an important point of asymmetry in my research. I wanted to analyze teachers' attitudes toward elected representatives linked to their union and whether the union contacted them during political campaigns. I obtained data from a 2009 teacher survey in Bogota, Colombia by Rocio Londoño and Javier Saenz (Londoño et al. 2011). Although this data shed light on the variables associated with teachers' political support for union leaders who entered public office in Bogota, I was unable to get comparable data for Argentina or Mexico. In my final analysis, I need to decide how prominently to feature this data because my research question concerns teachers in all three cases.

PROBLEM SOLVING

What should analysts do when they encounter roadblocks in the field? Data problems do not have easy, one-size-fits-all solutions. However, while frustrating, data problems are rarely insurmountable. Many workarounds require improvisation; they must be figured out on the ground, using limited information and gut feeling. This section considers strategies for coping with the aforementioned problems, and it considers guidelines on how to invest scarce resources.

Networks

Fieldwork is a messy, social process that involves building trust and establishing long-term relationships. It involves constructing networks that are multinodal; they should include a variety of different types of data brokers. This section sketches out some of the data brokers who can help access hidden data, confirm concerns that a database does not exist, and suggest strategies for efficiently pursuing unstructured data.

Gatekeepers can unlock hidden data. They can also allay concerns about whether data exists. Gatekeepers can be low-level functionaries, such as receptionists and assistants, who may make discretionary decisions about whether to help a researcher. I encountered gatekeepers who opened doors, and helped me obtain hidden data. These included sympathetic union leaders, government officials, and workers at nongovernmental organizations. I convinced these brokers to help me by cultivating relations of trust and conveying the important implications of my research for their organization. I also encountered gatekeepers who closed doors and prevented me from accessing data. Developing good working relationships with gatekeepers can yield unexpected positive consequences.

Organic intellectuals are nonacademic observers who know a lot about a given substantive topic. They may be analysts at think tanks, journalists, or public intellectuals. Organic intellectuals have tremendous substantive knowledge and can provide another perspective on a given research topic; they often interact with local academics. These brokers can help to correctly interpret data and understand how it was generated. In my research, organic intellectuals from the Colombian think tank the Center for Investigation and Popular Education, most notably Alvaro Delgado, gave me invaluable feedback as I advanced my project. Aside from having extensive networks and knowledge of how to get data, they also helped guide me when I had to make decisions about how to pursue unstructured data.

Fieldwork is a messy, social process that involves building trust and establishing long-term relationships. It involves constructing networks that are multinodal; they should include a variety of different types of data brokers.

Local academics are trained and work at local universities. They offer important insights regarding the substantive topic and usually have established professional networks. In Mexico, several distinguished local academics helped me connect to leaders in the teachers' union and gain access to hidden data. Karla Fernández, Aurora Loyo, Aldo Muñoz, and Carlos Ornelas all had established networks and access to valuable data sources that are not public. Local academics had a different perspective on my research question than professors at my home university. Theses Mexican professors knew substantially more about the Mexican teachers' union than my dissertation committee and their input kept the project on track during periods of confusion.

Veteran field researchers are usually trained in American universities and work on the same substantive topic. They have established networks and can also provide useful advice regarding data sources that can be accessed. Unfortunately, over time these academics may switch topics and lose touch

https://doi.org/10.1017/S1049096514000304 Published online by Cambridge University Press

with the data brokers they once knew well. For me, Maria Lorena Cook of Cornell University and Maria Victoria Murillo of Columbia University both generously shared their experiences studying teachers' unions and put me in touch with contacts, who offered to share data.

The importance of networks that include multiple types of data brokers reinforces arguments for an area studies approach to comparative research. Because it is costly to assemble these networks, the best way to organize research is to have analysts commit to studying a few countries. They can then develop relationships of trust and reciprocity, as well as deep case knowledge.

One question that should constantly be asked is, "is this data essential for this project?" Flexibility is important. Theory can be tested using nondata intensive methods.

Sequencing

Research often involves collecting a lot of data and then only using a small subset. The more a project is properly sequenced, the more time can be spent chasing critical data. Unfortunately, knowing what is most important for a given research question often can only emerge from the fieldwork experience itself. Properly sequencing a project can help to minimize wasted resources when putting together unstructured data or seeking out hidden data. Analysts must build theory before testing theory. Too often, because of skewed incentives, theory building is not given enough time. Grants are awarded for concrete data collection proposals—even if these proposals lack well-developed research questions. Pressure to publish induces a rush to collect data and test hypotheses. By moving forward too quickly, many well-funded projects end up wasting grant money. Theory building should involve iteratively moving between data and theory, and reflexively considering new directions for the project, based on what data is found to be available.

Researchers should invest in costly data collection activities at the end of a project. They should identify multiple data sources before deciding which to use. When the project is quite advanced, specific and costly tasks can begin: launching a survey, conducting content analysis of newspaper archives, or coding archival materials. It can be useful to perform triage, and differentiate between essential and nonessential (albeit quite interesting) data. It can also be helpful to decide whether data can be gathered in small pieces or whether it must be collected in one big chunk.

Sequencing can also be thought of in terms of an academic career. Graduate students have relatively few resources—in terms of research funds, research assistants, networks, and coauthors—in comparison to professors. Graduate students should be realistic about their constraints. They can plan to pursue bigger projects with more costly data collection activities later in their careers.

I carried out several short trips to the field before spending a full year conducting field research on teachers' unions. I dedicated a substantial amount of time to soaking and poking, experimenting with ideas, and trying to understand how teachers' unions actually operated. When my project finally developed a strong foundation, I began to carry out repetitive tasks—constructing a candidate recruitment database, searching for teacher survey data, and structured queries of newspaper archives.

One question that should constantly be asked is, "is this data essential for this project?" Flexibility is important. Theory can be tested using nondata intensive methods. For example, the concept of observable implications does not require "smoking gun" evidence in support of a theory (Collier 2011, 825–26). Instead, it requires limited, albeit significant, observations to bolster a theory's plausibility. A well-specified theory can be built around a thoughtful set of observable implications. It is not always necessary to laboriously compile a complete, rectangular dataset. Case studies are regularly used to test key assumptions and predictions of formal models. Some—although by no means all—projects can be defended with limited data.

Flexibility also involves thinking creatively about the substitutability of indicators and clever proxy variables. The concept of triangulation, now a cornerstone of multimethod research, can provide a path forward (Brady and Collier 2010).[7] This method enables analysts to make inferences based on multiple data sources—if none individually provide "smoking gun" evidence.

CONCLUSIONS

Data collection has important downstream consequences for data management and analysis. Too often, data collection is approached as an informal and idiosyncratic part of the research process and graduate students are expected to figure this out alone. This article aims to flesh out some common challenges that are encountered in the field and guidelines for thinking about how to confront them.

Fieldwork involves shoe leather. The epidemiologist John Snow, in dogged pursuit of data that would help him understand the spread of cholera, wore out the soles of his shoes walking around London (Freedman 2009). For political scientists, data collection involves going out into the street, talking to people, and trying to understand what is actually happening on the ground. Full immersion in the field is the best way to figure out the most interesting part of a research project and to find the data necessary to successfully complete it. ■

NOTES

1. Nurit Martínez. "Impulsa el SNTE a casi 2 mil candidatos." *El Universal* June 27, 2012.
2. Geraldo is a pseudonym.
3. For example, Eggers and Hainmueller (2009, 517-19) used a very costly data collection strategy in their analysis of the economic returns to office holding. They identified British MP candidates who barely won and barely lost elections by coding parliamentary candidate biographies from an archive of *The Times of London*. They then identified the date of death of these candidates, using a genealogy database, and collected data on their probate values (which is a legal record of the size of an individual's estate). This is a good example of a costly and complex data collection process that yielded substantial professional and analytic payoffs.

https://doi.org/10.1017/S1049096514000304 Published online by Cambridge University Press

4. Some activities involve a significant fixed cost, in the form of specialized training, but have a low marginal cost. For example, web scraping involves learning how to write code in a programming language. However, when an analyst can write code, the marginal cost of assembling data drops significantly. Amazon's Mechanical Turk and online surveys also have low marginal costs. Because these tools make data collection less costly, analysts can experiment with various datasets before committing to the one that yields the most interesting analysis.
5. By contrast, gathering data for theory testing or replicating the results of an existing study is focused and highly structured.
6. Chhibber and Kollman (1998, 338) acknowledge a problem of asymmetric data but are able to proceed in their comparison of the United States and India. In Table 1 they run a regression on the effect of government fiscal centralization on the effective number of parties in the United States. However, they are unable to find corresponding data in India to run the same regression. "Because there are too few cases for meaningful regression analyses, we do not replicate Table 1 for India."
7. Brady and Collier (2004, 310) describe triangulation as a "research procedure that employs empirical evidence derived from more than one method or from more than one type of data. Triangulation strengthens the validity of both descriptive and casual inference."

REFERENCES

Brady, Henry, and David Collier (editors). 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Oxford: Rowman and Littlefield.

Chhibber, Pradeep, and Ken Kollman. 1998. "Party Aggregation and the Number of Parties in India and the United States." *American Political Science Review* 92(2): 329–42.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science and Politics* 44(4): 823–30.

Eggers, Andrew, and Jens Hainmueller. 2009. "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103(4): 513.

Freedman, David A. 2008. "On Types of Scientific Enquiry: The Role of Qualitative Reasoning." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier (editors). *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press: 309–18.

Londoño, Rocio, Javier Saenz. Carlos Lanziano, Bibana Castro, Vladimir Ariza, and Mario Aguirre. 2011. *Perfil de los docentes del sector público de Bogotá*. IDEP, Secretaria de Educación Distrital.

https://doi.org/10.1017/S1049096514000304 Published online by Cambridge University Press