

Getting Started in Empirical Research

Teresa A. Sullivan, Executive Vice Chancellor for Academic Affairs, The University of Texas at Austin; Professor of Sociology and Law, The University of Texas at Austin¹

Empirical work is expensive, time-consuming, and hard. Empirical work requires training that relatively few law professors have had. But empirical work remains attractive and is, in at least some venues, influential. The attraction of empirical research is really two-fold. First, an empirical approach can take work from the realm of the normative – how the thing should work – to a frank look at how the thing really does work in the real world with all of its complexity. This is the descriptive function of empirical work, and in many areas of commercial law even the descriptive function would represent an advance. Secondly, an empirical approach can develop the data needed to test a hypothesis about how things *might* work under different conditions. Thus, Culhane and White were able to examine the effects of the then-proposed means test for Chapter 7 bankruptcy petitioners. Marianne B. Culhane and Michaela M. White, *Debt After Discharge: An Empirical Study of Reaffirmation*, 73 Am. Bankr. L. J. 709 (1999). When theoretical development is sufficiently robust, then hypotheses may be derived from the theory and tested using empirical data.

Primary versus Secondary Data

“Empirical” is a term that includes many methods, including both quantitative and qualitative approaches. Anyone considering an empirical project should consider first whether to use primary or secondary data. Primary data are those collected by an investigator for a specific project. I will discuss primary data in greater detail below.

Before making a commitment to collect data, a researcher would be well-advised to spend at least some time examining what is already available. “Secondary data” is a term that refers to data and databases that already exist and can be re-analyzed by investigators to address their specific research questions. Sometimes this is referred to as an “available data” approach. Archives, governmental information of all sorts, and the reanalysis of primary data collected by others are often used in secondary analyses. In both Canada and the United States, extensive data are collected by governmental agencies and are available to the public, often at little or no cost.

There is a substantial cost advantage to doing secondary analysis. Others have borne the cost of collecting the data, and although some costs may be incurred to tabulate the data, these costs will be considerably lower than original data collection. These existing data bases are often far superior in their extent, both geographic and historical, to anything that a single investigator could achieve. Most are in the public domain and many are easily accessible in a machine-readable format. Their major disadvantage is that the data were not collected to answer the investigator’s specific research questions.

¹ Mail address: Office of Academic Affairs, 601 Colorado Street, Suite 305, Austin TX 78701 U.S.A
tsullivan@utsystem.edu

A particular cross-tabulation or table may be impossible to calculate, or a key concept may need to be measured using second-best variables. (It is also worth noting that many important concepts are difficult to measure regardless of the data being used. An example in the current bankruptcy debates is the concept “stigma.”)

Some examples of data bases relevant to commercial law researchers include the Federal Reserve Board’s Survey of Consumer Finances (<http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html>, last consulted May 8, 2005) and the Economic Census of the United States (<http://www.census.gov/econ/census02/>, last consulted May 8, 2005) The Survey of Consumer Finances, conducted every three years, provides data on the balance sheet, pension, income, and other demographic characteristics of U.S. families. The survey also gathers information on families’ use of financial institutions and financial vehicles. The Economic Census of the United States, most recently conducted in 2002, is a somewhat under-used resource that provides detailed information on firms and industries with a variety of geographic specification. Statistics Canada provides searchable data on the Canadian economy and population (<http://www40.statcan.ca/101/cst01/> (English), <http://www40.statcan.ca/102/cst01/> (français), last consulted May 8, 2005). It is important to emphasize here that although many of the extant data bases use individual persons as the unit of analysis, there are many others that are organized by business or governmental entities and have rarely been mined by legal scholars.

There are also consumer surveys of many types that are publicly available at little cost. One extensive international archive is the Inter-University Consortium for Political and Social Research (ICPSR) (<http://www.icpsr.umich.edu/>, last consulted May 8, 2005). A query for the word “debt” at this site revealed 281 data sources, beginning with the World Debt Tables series, information on debt that was developed by the World Bank from the Debtor Reporting System (DRS) of the bank. Other data sources that are listed include many one-time surveys done by major news organizations on issues dealing with credit cards and debt. This site contains a full archive of the General Social Survey, a survey of a cross-section of the English-speaking adult population of the United States that includes a wide range of social, economic, and political questions.

In addition, data that have been collected by the U.S. federal government or through a grant from a federal agency are typically declared to be the public domain and made available to other researchers. A small cost for the actual reproduction of the data may be the only expense, although many of the data sets are now in electronic format and may be transmitted or shared electronically.

Another type of secondary analysis that can be done is meta-analysis. Meta-analysis is re-analysis of a number of published research studies to draw a conclusion. If there are, for example, a number of district-level studies of a legal process of interest, and if these studies are sufficiently comparable, then a meta-analysis could be used to generalize a larger conclusion. Meta-analysis is by now a fairly sophisticated technique of its own, but it shares with other secondary techniques the advantage of being relatively inexpensive. Meta-analysis can also make a substantial contribution to the understanding

of the subject if it is done credibly. For an introduction to meta-analysis, see Thomas D. Cook, *Meta-analysis for Explanation: A Casebook* (1992) and Gene V. Glass, *Meta-analysis in Social Research* (1981).

If none of the extant data fits the research question, however, the investigator should consider primary data collection. The great advantage of primary data is that the investigator can tailor the data collection to answer exactly the planned set of questions. The disadvantage is that primary data collection is so expensive and time-consuming that the studies are likely to be small and subject to the charge of being non-representative.

Primary data collection requires the investigator, or more commonly the investigative team, to undertake the identification, selection, storage, and systematization of the data. In these collection processes, great attention should be paid to data quality, particularly to any biases that might be inherent to the identification or collection of the data. Data that are missing, incomplete, or inaccurate can result in a flawed study with conclusions that are unreliable.² The expense of data collection, in both time and money, derives in large part from these considerations. In the next section, I address some of the issues of primary data collection in greater detail.

Collecting and Using Primary Data

Primary methods may be qualitative, such as interviews of key informants. This technique was used successfully by Ronald Mann to discover why the Japanese use of credit cards lagged so far behind the use in other countries with equivalent levels of industrialization. Ronald J. Mann, *Credit Cards and Debit Cards in the United States and Japan*, 55 VAND. L. REV. 1055 (2002). The use of semi-structured interviews with a number of respondents offers the opportunity to draw generalizations while still being able to understand the unique aspects of each respondent's situation. Deborah Thorne, for example, used this approach to understand the situation of a sample of couples interviewed a few months following their bankruptcy discharge. Deborah K. Thorne, *Personal Bankruptcy Through the Eyes of the Stigmatized: Insight into Issues of shame, Gender, and Marital Discord*, unpublished Ph.D. dissertation, Washington State University (2001). Compare this study with an interview study of a smaller group of individual debtors. Sheila Jo Wojcik, *The Post-Bankruptcy Rebuilding Process: How the Chapter 7 Debtor learns to Begin Life Anew*, unpublished Ph.D. dissertation, The University of Texas at Austin (2002). A generally useful critical approach to interviewing is Elliot G. Mishler, *Research Interviewing: Context and Narrative* (1986).

Even for a project that will subsequently be more quantitative in direction, interviews with key informants can be very helpful in shaping the subsequent study. For example, a series of semi-structured interviews with bankruptcy judges and attorneys proved to be key in understanding why there were so many local variations in bankruptcy, despite a common federal law governing the bankruptcy process. This interview information was critical in the subsequent data collection and data analysis of a

² There are statistical measures for correcting for some of these flaws, but it is preferable to design the data collection to minimize the quality issues as much as possible.

data base of bankrupt debtors. Teresa A. Sullivan, Elizabeth Warren, and Jay Lawrence Westbrook, *The Persistence of Local Legal Culture: Twenty Years of Evidence from the Federal Bankruptcy Courts*. 17 HARVARD J. OF LAW & PUBLIC POLICY 801 (1994); Teresa A. Sullivan, Elizabeth Warren, and Jay L. Westbrook. *Consumer Bankruptcy in the United States: A Study of Alleged Abuse and of Local Legal Culture*. 20 J. OF CONSUMER POLICY 223 (1997).

Content analysis is a qualitative method that is customarily used with documents, but may also be used to analyze transcripts. Content analysis software is inexpensive and readily available, and allows the analyst to search for key phrases and words in their context, and then to draw conclusions based on their relative frequency and importance. Content analysis is useful for identifying themes and patterns in written materials. A classic example of content analysis (albeit with some statistical treatments included) is Frederick Mosteller and David L. Wallace, "Deciding Authorship," in J. Tanur et al., eds., *Statistics: A Guide to the Unknown* 164 (1972), an effort to identify the authors of specific *Federalist* papers by their writing style.

Quantitative methods, including those involving sophisticated statistical modeling of the study in question, are appropriate for some types of data. Before a quantitative analysis can be undertaken, however, the primary data must be collected in such a way that they are amenable to statistical manipulation. Ideally, the anticipated analysis will guide the collection of the data. At a minimum, the investigator must know which concepts will be critical to measure.

Any quantitative approach requires that the information for each case be reduced to variables, each of which is intended to measure a concept. To take a simple case, the concept "age" can be captured by a variable that asks for current age, age at last birthday, or date of birth. It is a bit more of a stretch, but still reasonable to calculate age based on date of high school graduation. It would be much more dubious, and would raise more doubts about the validity of the measure, to measure age by date of law school graduation or by the date of tenure in a university appointment. In designing a study, if age is to be a variable of interest, then the investigator needs to consider how best to assess age. There is a substantial literature on this subject; in many populations, for example, older people round off their ages to multiples of five, so that a question asking directly for age may be less accurate than a question that asks for date of birth. I present this relatively simple question by way of illustration, recognizing that the issues addressed in commercial law are much more complex. A question such as "What is the current value of your inventory?" may seem easy to answer at first blush and yet involve many issues such as depreciation, inflation, and so on. And unlike the basic demographic questions such as age, there is less likely to be a helpful body of literature to offer best practices.

In general, statistical approaches can establish correlations or associations among variables. Causality is more difficult to establish, even when there is a known time relationship among variables. Statistical approaches can also offer guidance as to whether a relationship between variables has come about because of chance variation, or whether the relationship is "significant" – that is, unlikely to have come about through

chance. Most importantly, multivariate statistical analyses offer the possibility to establish significant relationships even though many variables may be involved. A multivariate technique allows the investigator to “hold constant” the effect of variables that might confound the relationship. Multivariate analyses thus allow investigators to identify the most important contributors to a relationship when several variables might be compete. Moreover, a multivariate method can identify the interactions of variables – that is, when several factors multiply the strength of a relationship. For example, a full moon, a high tide, and warm temperatures will intensify the effects of a hurricane, other things equal. An example of such an analysis is Teresa A. Sullivan, Elizabeth Warren, and Jay Lawrence Westbrook, *Who Uses Chapter 13?* Pp. 269-282 in Johanna Niemi-Kiesilainen, Iain Ramsay, and William C. Whitford, eds. *Consumer Bankruptcy in Global Perspective* (2003).

Most statistical methods, however, assume that sampling techniques have been used. A problem with many empirical studies is the difficulty in getting a representative, let alone random, sample during the data collection. In commercial law studies, it might be more interesting to study the non-representative cases – for example, firms with the highest dollar value of assets, largest number of employees, greatest indebtedness, and so on. I have previously elaborated on the sampling problem. See Teresa A. Sullivan, *Methodological Realities: Social Science Methods and Business Reorganizations*, 72 WASH. U. L. Q 1291 (1994).

Appropriate modesty

Given the great difficulty of using true experimental approaches to the issues studied by the law and social science community, there are few studies that can be said to be definitive in the sense that they have controlled for every confounding variable. Good empirical work is usually painstaking in noting ways that the data or methods may be defective. Alternative analyses, alternative explanations, and attempts to replicate are all part of the intellectual process of assessing the value of the empirical research. In the end, data do not “prove” anything, although they may serve to disprove a prevailing hypothesis or shift the grounds of debate. Proprietary data that cannot be further examined or studies with restricted access are problematic at best. For further elaboration, along with additional caveats for the neophyte to empirical research, see Jay Lawrence Westbrook, *Empirical Research in Consumer Bankruptcy*, 80 Tex. L. Rev. 2123