Gathering Data for Financial Reporting Research

Robert Bloomfield
S.C. Johnson Graduate School of Management
Cornell University

Mark W. Nelson
S.C. Johnson Graduate School of Management
Cornell University

Eugene Soltes
Harvard Business School
Harvard University

April 4, 2015

This paper has been prepared for the 2015 Journal of Accounting Research Conference. We thank Matthew Bloomfield, Scott Emett, Ryan Guggenmos, Anne Heinrichs, Scott Jackson, Robert Libby, Karthik Ramanna, Doug Skinner, Jennifer Tucker, Isabel Wang, and Anastasia Zakolyukina for comments, and Rajesh Vijayaraghavan for his assistance gathering data on published accounting research.
Gathering Data for Financial Reporting Research

Abstract

In the published proceedings of the first Journal of Accounting Research Conference, Vatter [1966] lamented that “Gathering direct and original facts is a tedious and difficult task, and it is not surprising that such work is avoided.” For the fiftieth JAR Conference, we introduce a framework to help researchers understand the complementary value of seven empirical methods that gather data in different ways: prestructured archives, unstructured (“hand-collected”) archives, field studies, field experiments, surveys, laboratory studies, and laboratory experiments. The framework spells out five goals of an empirical literature and defines the seven methods according to researchers’ choices with respect to five data gathering tasks. We use the framework and examples of successful research studies to clarify how data gathering choices affect a study’s ability to achieve its goals, and conclude by showing how the complementary nature of different methods allows researchers to build a literature more effectively than they could with less diverse approaches to gathering data.
Gathering Data for Financial Reporting Research

1. Introduction

To open the published proceedings of the first Journal of Accounting Research Conference (May, 1966), Sidney Davidson wrote:

Accounting is singularly concerned with the quantitative expression of economic phenomena. Despite this preoccupation with numerical presentation, there has been little empirical analysis or testing of the concepts of accounting. Accounting thought will develop more effectively by increased reliance on the testing of meaningful hypotheses; there is a need to look to evidence as well as to authority for the substantiation of accounting ideas (Davidson [1966]).

Fifty years after the first JAR conference, most accounting studies fulfill Davidson’s exhortation to “look to evidence,” and in so doing have allowed researchers to understand and predict the causes and consequences of many accounting phenomena. However, the literature still struggles to confront the challenge laid out in William Vatter’s “Critical Synthesis of Conference Papers”, which closed the published proceedings:

One of the real limitations of empirical research is that we tend to work on problems we are able to study because data are available; we thereby tend to overlook problems that we ought to study, if data for such problems are not easy to obtain. It is significant that the larger and more comprehensive efforts reported here have dealt with published or otherwise readily available data. Gathering direct and original facts is a tedious and difficult task, and it is not surprising that such work is avoided (Vatter [1966]).

Our goal in writing this paper is to help accounting researchers realize Vatter’s vision over the next 50 years, much as they have realized Davidson’s in the last 50. We know that “gathering direct and original facts is a tedious and difficult task”, and we cannot make it any easier. But we can help researchers gather data wisely and explain their choices and contributions to readers, reviewers, and editors. We do so by introducing a framework that spells out five goals of an empirical literature and five data gathering tasks that researchers can use to advance those goals. We use the framework to define seven methods that appear in empirical accounting research, to clarify how data gathering choices affect a study’s contribution, to recommend some practices that will enhance that contribution, and to show how a community of scholars build a literature more effectively by using a wider range of methods.
The paper proceeds as follows. In Section 2, we describe five goals that an empirical literature seeks to accomplish: (1) specifying causal theories to test, (2) testing for predicted associations between variables, (3) attributing those associations to the causal theories, (4) verifying robustness and generality of results, and (5) placing results in context and offering additional opportunities for theory building. A successful literature, taken as a whole, strives to achieve all of these goals. Any particular study is likely to emphasize some goals more than others.

In Section 3, we identify five data-gathering tasks that researchers either can choose to undertake or delegate to others (or in some cases, nature). The first two tasks help researchers distill observations into variables that are well-suited to test their causal theory. Researchers can choose whether to (1) record observations specific to testing their theory or use records that were created by others for more generic purposes, and (2) hand-collect records to convert them into structured data sets amenable to statistical analysis or use data prestructured by others. The other three tasks involve researchers intervening in the data-generating process to record data to test their theory. Researcher can choose whether to (3) elicit dependent variables or observe those variables, (4) manipulate independent variables or allow variation to arise naturally, and (5) control other variation in the setting or allow that setting to vary naturally. In Section 3.1, we discuss how choices with respect to each of these tasks can help a study achieve some goals at the expense of others. In section 3.2, we define seven distinct methods according to the bundle of data gathering tasks that the researcher chooses to undertake. We derive the seven methods by assuming that two studies use the same method if the researcher undertakes the same set of tasks; otherwise, the studies use different methods.

In Section 4, we discuss a number of the framework’s implications. The framework draws several useful distinctions between methods, indicating two distinct types of archival study (depending on whether the researcher hand-collects data or uses a prestructured archive), two forms of laboratory study (depending on whether the researcher manipulates an independent variable), and a narrow definition of field study (because it requires that the researcher record original data). The framework also indicates that many studies (or parts of studies) that are called surveys actually apply a laboratory study or laboratory experiment method. The framework does not provide a basis for recommending one method over another based on general factors like the source of theory (economic, psychological) or the type of behavior
investigated (individual, group, market). Such choices are driven by the particular circumstances in which the study is conducted: the state of theory and prior empirical findings, the availability of data archives and technology for hand-collection and intervention, and the characteristics of naturally occurring phenomena.

In Section 5, we provide recommendations that researchers can follow when applying each method. Because recommendations are so strongly influenced by the particular circumstances of a study, we do not derive them from the framework. Instead, we identify them by referencing studies that illustrate wise choices that exploit the advantages and overcome the challenges of the chosen method to achieve research goals.

In Section 6, we use extant research on recognition versus disclosure to illustrate how an empirical literature can thrive by including complementary contributions from a variety of methods to advance our knowledge of important accounting phenomena. In Section 7, we provide brief closing comments.

To keep the paper a manageable length, we restrict our attention to studies that extend the financial reporting literature by helping to confirm or refute theories used to explain observed behavior. We define financial reporting research broadly to include investigations of all causes and consequences of reporting financial information to external stakeholders, and the mechanisms by which those causes and consequences arise. We do not intend this paper to be a comprehensive review of any literature, but rather draw examples from financial reporting to illustrate key points. By focusing on financial reporting, we exclude other important accounting literatures, but hope that readers in other areas will be able to apply our framework to their own work. By focusing on empirical papers that test theory, we exclude many valuable forms of theoretical and qualitative research common in the social sciences (such as economic modeling, history, and interpretive research), but hope that those who conduct such research will find useful insights that help them draw from and speak to theory testing. By focusing on select examples that illustrate how to use our framework to inform research choices, we also exclude many excellent papers even within financial reporting.

2. Five Goals of Empirical Literatures

An empirical literature strives to accomplish five goals: (1) specifying the theory being tested, (2) testing for statistical associations between variables, (3) attributing those associations
to the causes indicated by the theory being tested, (4) verifying that the association and its attribution generalize across variations in execution, data and setting, and (5) placing results in context and offering additional opportunities for theory building. Any individual study typically accomplishes some goals more effectively than others. Its contribution depends on the extent to which it helps the literature as a whole achieve these same goals. We elaborate on each of these goals according to the sub-goals and definitions in Table 1.

2.1 SPECIFICATION

We use a causal map to clarify our view of theory specification. We define theory as an assertion that two or more constructs are causally related. The simplest possible prediction is that variation in construct A causes variation in construct B for the reason specified by the theory, as shown by the boxes of Figure 1. In the simple example shown in Figure 1, the causal theory is that higher reporting quality (construct A) causes firms to enjoy a lower cost of capital (construct B) because capital providers consider those reports to be more useful for decision making. As theories mature, additional studies contribute by extending or clarifying constructs. One can extend the constructs by hypothesizing that reporting quality is caused by corporate governance (adding a distal cause), that the firm’s cost of capital affects its ability to generate value through investments (adding a distal effect), that reporting quality affects cost of capital through its effect on market liquidity (adding a mediating construct), or that cost of capital is more sensitive to reporting quality for firms with high transactions costs (adding a moderating construct). One can clarify constructs by distinguishing between different aspects of reporting quality (such as freedom from manipulation, conservatism and timeliness) and showing that these more distinct constructs have different effects.

2.2 ASSOCIATION

Constructs like “reporting quality” and “cost of capital” are abstract concepts that are claimed to have a causal relationship. Because abstract concepts are not directly observable, researchers create a measurable proxy for each one, and infer the causal relationship between

---

1 These five goals bear important similarities to Kuhn’s (1969) definition of the paradigm embraced by a scientific community, which includes key constructs, modeled causal relationships, shared agreement regarding what it means for a prediction to be accurate, types of evidence considered relevant and persuasive, and types of predictions considered important and useful.
constructs from the statistical association between their proxy measures. We represent these relationships visually using Libby Boxes (see Libby [1981] and Kinney and Libby [2002], drawn from Runkel and McGrath [1972]). As shown in Figure 2, the theory predicts that reporting quality has a causal effect on the cost of capital (link 1). The research design operationalizes the independent variable “reporting quality” by measuring deviations in accruals from what one would expect absent intentional manipulation (link 2), and operationalizes the dependent variable “cost of capital” as the firm’s average stock return (link 3). Statistical tests establish the association between the operationalized independent and dependent variables (link 4). Demonstrating a convincing association between these variables requires accounting for other variables that might affect the dependent variable but are omitted from the theory (link 5). Researchers make tests of associations convincing by maximizing power, minimizing noise and bias, and reporting the data and analysis clearly enough to allow the reported associations to be verified by others.

2.3 Attribution

Having identified an association, researchers then seek to attribute the association to the causal factors they have specified. Such attributions become more effective as they strengthen the links connecting their Libby boxes. The more precise the specification of the causal theory (link 1), the tighter the connection between constructs and their operationalizations (links 2 and 3), the more reliable the tests of association (link 4), and the more comprehensively the design controls for omitted variables (link 5), the more likely the statistical association is attributable to the specified causal relationship.

Attribution requires establishing four facts: (1) causation runs in the stated direction (in our example, the association isn’t driven by firms with higher returns choosing to report greater unexpected accruals); (2) the association is driven by an appropriately narrow definition of the constructs specified (e.g., only the discretionary component of reporting quality drives returns); (3) the association is mediated and moderated as theory would predict (e.g., for mediation, unexpected accruals reduce liquidity, which in turn increases returns; for moderation, the association of unexpected accruals with returns is greater when investors are more risk averse); and (4) the association isn’t driven by variables other than the specified cause (such as firm size or age).
2.4 Generalization

Even if a literature has successfully attributed an association to a causal theory, it must then provide reason to believe that interpretation will generalize beyond the specifics of each study. We distinguish between four types of generality. The first three are relatively straightforward. Robustness demonstrates that the results are not dependent on specific measures or analysis techniques, and is achieved when researchers produce similar results using new techniques with the same data. Replication demonstrates that the results are not dependent on a particular data set, and is achieved when researchers produce similar results using similar techniques with new data. Triangulation demonstrates that the results are not dependent on a particular method, and is achieved when researchers produce similar results using different methods for gathering data. As discussed below, different methods have different advantages to exploit and different challenges to overcome, so triangulation provides evidence that results are not specific to one method.

If results hold up to robustness checks, replication and triangulation, the literature should have relatively high confidence that the results provide accurate attributions about the causal theory being tested. Returning to our example, if an association between reporting quality and cost of capital holds up to (1) different measures of reporting quality and cost of capital, (2) is evident in different samples of data, and (3) is evident in archival data, lab experiments, and surveys, we should have high confidence that reporting quality affects cost of capital.

A fourth type of generalization is application. Application occurs when actions are taken in light of the theory within practice settings. Application can be viewed as the ultimate goal of an applied social science like accounting: using what we have learned to predict future behavior of the financial reporting settings that motivate the research, and using those predictions to propose actions that will improve the welfare of actors and/or society. Returning to our example, application would occur when managers, regulators or standard setters use research to guide their efforts to reduce the cost of capital by improving reporting quality.

Individual studies typically provide only limited empirical demonstrations of generality. Most demonstrate some degree of robustness (showing that the results are similar using different

---

2 Replication is sometimes confused with recalculation, which occurs when a researcher accesses the same data and performs the same tests as in the original study. Recalculation is a useful technique for ensuring that, e.g., an archival study has accurately reported its sample-selection criteria and statistical analysis, but it does not provide assurance that the same results would be produced by a different data set.
statistical techniques), and some demonstrate a degree of replication (showing that the results are similar across multiple data sets). Most rely on subsequent work for replication, triangulation or application.

2.5 CONTEXTUALIZATION

Researchers place their tests in context by reporting findings that allow readers to understand the economic importance of their results and the behaviors and institutions in which the data were generated. Contextualization is facilitated by providing descriptive data, reporting estimates of economic significance, and conducting exploratory analyses that provide information about the context that was not theorized previously to exist. Returning to our example, one might estimate that a manageable improvement in reporting quality reduces a firm’s cost of capital by 10%, report an unexpected association between reporting quality and the educational background of executive officers, or list particular tools managers use to affect reporting quality. Such findings can guide future empirical work by suggesting new theories to be tested and new data to gather.

3. Data Gathering and Research Methods

Researchers use a variety of methods to achieve the five goals described in Section 2. Table 2 shows the relative prevalence of five categories of methods used for various topics in the 1,271 papers published in the Journal of Accounting Research, the Accounting Review and the Journal of Accounting and Economics from 2003 to 2013.³ To classify each paper, we relied on the authors’ description of the primary data gathering method used to test the hypotheses in the paper. When papers employed a non-archival method of data gathering (alone or in addition to archival data), we sought to classify the paper according to that non-archival choice. Still, the vast majority of papers rely on archival (historical) data, with many papers using pre-existing archives (like CRSP and Compustat), but some developing archives by hand-collecting documents (like the full text of annual reports or equity analyst reports).⁴ We find that only a tiny fraction of published papers are based primarily on data gathered through field studies and

³ For example, The Accounting Review’s 2014 editorial report categorizes empirical papers as empirical-archival, experimental, field and case study, survey, and other (Evans [2014, p. 2347]).
⁴ The archival group includes papers analyzing proprietary archival data, which, as discussed in section 4.2, can include a significant field-study component.
field experiments, which most directly pursue Vatter’s goal of “gathering direct and original facts” by capturing events that occur in practice settings in the ordinary course of business. More papers gather data from experiments and surveys, which are perhaps less direct, but still record original data.

While the list of methods in Table 2 is familiar, it provides little guidance to researchers who wish to efficiently and effectively engage in the “tedious and difficult task” of gathering new data. To provide that guidance, we decompose data gathering into five distinct tasks. In Section 3.1 we define each task and discuss how undertaking it can help a researcher achieve some of the research goals listed in Section 2, while potentially hindering others. In Section 3.2 we define different research methods by the bundle of tasks the researcher chooses to undertake.

3.1. DATA GATHERING TASKS: DISTILLATION AND INTERVENTION

Data gathering tasks fall into two categories, distillation and intervention. Distillation tasks convert observed phenomena into data that is designed to calculate statistical associations that can be attributed to the researcher’s theory. Intervention tasks alter the setting in which data are generated in ways the researcher believes will make causal attribution easier.

3.1.1 Distillation. In William James’ memorable phrase, babies experience a “buzzing, blooming confusion” (James [1890]) until they learn to organize their perceptions into a framework that allows them to distinguish objects, events, causes and effects. In the same way, a researcher must observe phenomena and organize those observations into variables that allow the researcher to identify statistical associations and attribute them to the theory they have specified. We call this process distillation, because it requires the researcher to extract the key information that allows them to operationalize the constructs included in their causal theory, as well as obtain other data that help them test associations and make causal attributions. Distillation requires two tasks: data recording and data structuring.

3.1.1.1 Data recording. The first step in distillation is to make records of observed phenomena. Researchers who undertake the task of data recording observe what occurs in practice settings and record those observations as numbers, text, images, audio, video and lists of facts and events that the researcher believes will be useful for theory testing. This contemporaneous observation helps the researcher understand both the data-generation process.
and the context to which the theory is intended to apply, but takes a great deal of time, and risks distorting the behavior of people who know they are being observed.5

Alternatively, the researcher can delegate data recording by relying on archives of records that have been created by someone else and made available through a data aggregator (such as Dow Jones, which provides Factiva, an archive of press releases and news articles), a regulatory body (like the SEC, which provides EDGAR, an archive of information contained in corporate reports), a company that creates records of its own activities in the ordinary course of business, or, more rarely, a prior researcher who has gathered survey, field or laboratory data. Researchers who delegate data recording benefit by gaining access to large archives that provide great power for tests of association and robustness. However, they also allow another party to organize the “buzzing, blooming confusion” of phenomena into records. That other party will choose what information to capture, how to capture it, and what information to ignore based on their own views, rather than tailoring those choices to the researchers’ needs.

3.1.1.2 Data structuring. The second step in distillation is to convert records of observations into data sets that are structured to allow statistical processing. While structures can vary, the most familiar consist of tables that include a row for every observation and a column for every variable. Every variable must be coded with numerical or categorical measures to permit statistical computation. Such coding typically involves a great deal of information loss through omission and aggregation. For example, consider the challenges of structuring an archive of text documents. Numerical variables for each document might include the number of words, the number of words per sentence, and the number of words within each of various categories (positive words, first person pronouns, references to competitors). Categorical variables might include the author, audience and topic. However, it would be impractical to create a variable for every dimension on which the text might vary, so much context is lost in the data-structuring process.

Researchers can structure data for themselves to ensure that variables are coded in a way that is most likely to capture the constructs specified in the theory they wish to test. However, the “difficult and tedious” nature of those structuring tasks typically means that the researchers

---

5 The idea that people behave differently when they know they are being observed is sometimes referred to as the “Hawthorne effect”. Although prior work disputes attribution of various results to the Hawthorne effect (Levitt and List [2011]), it remains a consideration for researchers who gather their own data by observing behavior in laboratory or field settings.
must restrict their coverage of the setting. These studies may only examine structured records for tens or hundreds of firms (or, in the case of field studies, only one firm). As a result, researchers who structure their own data typically have less power to demonstrate association.

Alternatively, researchers can delegate data structuring to a third party. For example, many accounting researchers delegate to Standard and Poor’s the structuring of documents filed with the Securities and Exchange Commission. Standard and Poor’s creates categorical and numerical variables for each filing based on what it believes will be useful to its customers. Economies of scale allow the resulting structured database, Compustat, to provide many researchers with enough observations and coded variables to demonstrate powerful associations for a wide range of theories. However, a researcher who delegates structuring in this way may not find a measure that is well tailored to the particular theory they wish to test, thus hindering attribution. The generic nature of the data also makes contextualization more difficult, since much of the richness of the setting has been filtered out.

3.1.2 Intervention. Much useful data is generated from practice settings in which all events arise in the ordinary course of business, without any intervention from the researcher. However, researchers often find it useful to intervene in the data-generating process to improve their ability to attribute associations to their theory. We identify three types of interventions: eliciting dependent variables, manipulating independent variables, and controlling the data-generation setting in other ways.

3.1.2.1 Eliciting dependent variables. Researchers elicit dependent variables when they pose questions and present tasks to people and observe responses that otherwise would not occur, as is the case in surveys, lab studies and lab experiments. Intelligently elicited dependent variables can more precisely capture the construct being examined than would data that arises naturally, facilitating association and attribution. However, elicitation can influence the data that is observed, potentially undermining attribution. Also, the elicited variable might differ in important ways from outcomes of interest in the practice setting, limiting generalization and contextualization.

3.1.2.2 Manipulating independent variables. Researchers manipulate independent variables by developing multiple versions (called treatments) of a task that differ only according to pre-specified values. This manipulation of independent variables is the defining feature of experiments (either field or laboratory). By randomly assigning participants to treatments and
examining associations between the independent and dependent variables, researchers can obtain particularly strong evidence regarding the causal relations underlying their theory. Also, random assignment distributes across treatment levels other features of participants that might affect associations, obviating the need to record and analyze data about those effects unless they are constructs within the theory being tested. However, manipulating independent variables requires that the researcher either alter key elements of the practice setting, which might be difficult to do, or construct versions of the practice setting that might differ from it in important ways, potentially limiting generalization and contextualization.

3.1.2.3 Controlling the data-generating setting. Researchers who perform laboratory studies or laboratory experiments can control the data-generating setting to reduce or eliminate the influence of factors that aren’t germane to testing their theory, and emphasize factors that are. For example, the researcher can hold important information constant, or abstract from practice settings to omit unimportant or confounding information. Similar to other interventions, controlling the data-generation setting facilitates association and attribution but potentially reduces generalization and contextualization.

3.2. Defining Methods

We define a method as a unique bundle of choices over which data gathering tasks to undertake and which to delegate to others (or to nature). These definitions are most accurate when applied to individual theory-testing exercises that gather data to answer a single research question. They are less descriptive of entire papers, researchers or literatures, all of which often use many methods. While five tasks allow 32 theoretically possible methods (2 x 2 x 2 x 2 x 2), the seven depicted in Table 3 capture virtually all empirical theory-testing studies we observe in the financial-reporting literature. Figure 3 depicts the methods in a decision tree that clarifies why some combinations are not observed. For example, if you delegate data structuring to another party, you cannot undertake any of the other tasks. We walk through the methods in the order they appear in the decision tree.

Prestructured archival studies use archives of naturally occurring data that have been recorded and structured by others. For example, the providers of CRSP, Compustat, and other data sell access to prestructured archival data. All data-gathering tasks have been delegated to the data providers, who (in most cases) themselves avoid intervening in the setting. The
researcher identifies the data within the preexisting structures that best capture the specific theoretical constructs being tested for association, recognizing that those structures were created for more generic purposes.

**Hand-collected archival studies** gather data by structuring records drawn from existing unstructured archives to create measures that match the constructs specified in their theory. The researcher delegates recording to others and avoids intervening in the data generating process. When a study includes a mix of hand-collected and prestructured archival data, we classify it as a hand-collected archival study. We use the term “hand-collected” metaphorically, not literally. Sorting through a box of documents and typing numbers and categories into a spreadsheet is a familiar means of imposing structure on unstructured archives, but researchers can accomplish the same goal by writing a computer program to process machine-readable documents containing text or images.

**Field studies** record observations of naturally-occurring settings without the researcher intervening in those settings. Thus, as shown in Figure 3, the researcher has chosen to observe a dependent variable rather than eliciting it, and has chosen not to manipulate independent variables or control the setting in other ways, instead delegating variation to nature. Field studies are therefore similar to archival studies, except that the researcher creates the archive of records.

**Field experiments** differ from field studies by the researcher intervening in one way – the researcher manipulates independent variables. Thus, field experiments observe dependent variables as data occurs naturally in practice settings, but their manipulation of an independent variable allows strong causal inferences concerning effects on dependent variables.

**Surveys** also involve researchers intervening in an otherwise natural setting in one way – the researcher elicits dependent variables and other information that might be used to measure independent variables, capture other information useful for theory testing, and allow exploratory analyses. Surveys are used to elicit practitioners’ opinions or their recalled anecdotes. Some studies use survey methods exclusively, but survey data also often augment data from other research methods, including field and laboratory studies and experiments. Likewise, laboratory studies and laboratory experiments as we define them are often embedded in surveys.

**Laboratory studies** intervene in two ways, by eliciting dependent variables and by controlling the setting in which data are observed, but do not manipulate independent variables. This approach is somewhat common in experimental economics, e.g., with participants in a
laboratory market interacting over time and the researcher measuring the outcomes of those interactions. We use the term “laboratory” metaphorically, not literally. Placing participants in a physical laboratory is a particularly effective way to control the data-gathering process, but researchers can impose similar (but typically weaker) controls when gathering data at other locations or soliciting responses by the internet, and we refer to these as “laboratory studies” as well.

**Laboratory experiments** intervene in all three ways, eliciting dependent variables, manipulating independent variables, and controlling the setting. Laboratory experiments typically create an experimental task that depicts key features of a real-world practice setting and examine how manipulating features of that setting affects the judgments and decisions of investors, analysts, auditors, board members and other participants in the financial reporting process.

4. **Implications of the Framework.**

Our approach in this paper is to define methods by first identifying the goals that empirical researchers seek to accomplish and the data gathering tasks they use to accomplish those goals. This deductive approach confirmed many of our prior intuitions, but yielded unexpected implications for how to distinguish between methods, how to think about internal and external validity, and when to recommend one method over another. We discuss each of these implications in turn.

4.1. **Distinguishing Between Methods**

Our deductive approach leads us to definitions of methods that are mostly familiar, with three unanticipated refinements. First, the framework yields a distinction between two categories of archival research. Researchers analyzing prestructured archives delegate all data gathering tasks to someone who collected the data for a purpose other than testing the researchers’ specific theory. In contrast, researchers analyzing hand-collected archives convert unstructured archives of text, images, video and historical events into structured archives of measured variables. While these methods usually are lumped together, the structuring inherent in hand-collecting archives allows researchers greater ability to customize their data to capture the constructs they wish to examine.
Second, the framework yields a narrow definition of a field study, because it requires that researchers performing a field study gather their own original records, like an anthropologist who “lives with and lives like the subjects being studied” (Van Maanen [2011, p. 2]). Many studies use fieldwork—close interaction with the subjects being studied—to gain access to proprietary data and insight into practice settings to guide their analysis. Fieldwork can be of tremendous value, especially when it brings new data to light, but does not provide the flexibility for customizing data that comes from creating records from direct observation of field behavior.

Third, many studies that are commonly referred to as surveys are categorized by our framework as either laboratory experiments or laboratory studies, because they elicit responses from participants who are presented with scenarios (with or without manipulation of independent variables) that are tailored to the researcher’s theory. The framework does not define surveys by how participants communicate their responses (e.g., by phone, email or in person), but rather according to whether elicited responses occur in a situation that arises naturally, is controlled by the researcher, or includes manipulated independent variables.

4.2. INSIGHTS INTO INTERNAL AND EXTERNAL VALIDITY

The framework also provides insight into the ubiquitous but misunderstood concepts of internal and external validity. As defined by Cook and Campbell (1979, p. 37), internal validity “refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause”, and external validity “refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings.” As Jimenez-Bueod and Miller (2010) point out, internal validity is often said to be a prerequisite of external validity, while at the same time internal validity is often said to require a sacrifice of external validity. Our framework accommodates both views by distinguishing between the goals of empirical research and the tasks involved in gathering data.

When viewed in terms of empirical research goals, internal validity is a prerequisite for external validity, because external validity requires the researcher to generalize support for the causal relations inherent in theory. The only way to generalize a causal relationship is to first attribute observed associations to a specified theory, having confidence that the results will be robust to minor changes in analysis and that the results will replicate when similar data are
gathered. Once the researcher achieves these goals, which we see as capturing internal validity as it is typically conceived, the researcher can demonstrate external validity: the attribution will be supported by triangulation when data are gathered with different methods, it will explain phenomena in the target setting, and it can be applied to that setting to achieve desired outcomes.

When viewed in terms of data-gathering tasks, choices that enhance internal validity may sacrifice external validity. For example, researchers can intervene in the target setting, or create a setting of their own, for the explicit purpose of attributing associations to their specified theory. Such interventions make causal attribution easier, enhancing internal validity, but may make it harder to generalize and contextualize results, because the researcher has to defend that the associations and therefore their attributions apply to the target setting that their theory is intended to explain.

4.3. RECOMMENDING ONE METHOD OVER ANOTHER

Finally, the framework clarifies why it is so hard to recommend one method over another based on broad factors, such as the source of theory (e.g., economics or psychology) or the level of aggregation (individual, group or market) of interest. Methods typically differ in more than one task, each of which affects multiple goals, so such comparisons require assessing numerous tradeoffs. Moreover, those tradeoffs depend on the state of theory and prior empirical findings, and the availability of data archives as well as technology for hand-collection and intervention.

We therefore make no recommendations about when researchers should choose one method over another, and caution readers to view such claims with a great deal of skepticism, instead evaluating each study’s choice of method in light of its particular goals and circumstances.

5. Data-Gathering Choices Within Methods

Once a researcher has chosen a method, they face many choices on exactly how they will execute the data gathering tasks before them and how they will work with data resulting from tasks they have delegated. For each method, we identify advantages researchers can exploit and challenges they must overcome, and make some recommendations on how best to do so. We illustrate our recommendations using examples of research studies that we believe have made wise choices.
5.1 PRESTRUCTURED AND HAND-COLLECTED ARCHIVAL STUDIES

Some of the most influential studies in accounting analyze data from preexisting structured archives that capture information generated in the ordinary course of business and without any researcher intervention.⁶ Those studies delegate all data recording and data structuring to third-party prestructured data providers such as Compustat (for financial data), CRSP (for market data), and I/B/E/S (for analyst forecasts). A growing number of studies supplement prestructured data with what we label “hand-collected” archives, in which the researcher structures unstructured records. Sometimes the records are provided by a data aggregator, such as Factiva or Lexis/Nexis. More recently, researchers have begun “scraping” data directly from the internet, culling data about social media activity, web traffic, search behavior, and even extracting information from pictures and audio. Our framework defines a study as analyzing a prestructured archive if it includes data structured entirely by third parties; if the researcher supplements the prestructured archive by coding additional variables from an unstructured archive, the study is classified as hand-collected.

Prestructured and hand-collected archives have very similar advantages to exploit and challenges to overcome, differ on only one task, and may differ only on a small number of variables, so we discuss them together. For both, someone else has performed the “tedious and difficult task” of recording a large amount of data and (in the case of prestructured archives) has structured it in a standardized format amenable for analysis. The resulting large volumes of data allow powerful tests of association, and the standardized formats and broad availability foster the formation of a research community that can establish the robustness of colleagues’ results, collaborate to find better ways to use the data, and train new scholars. Indeed, third-party data providers develop and maintain prestructured archives because there is demand for the data that they provide. Also, because most of the records in these archives are created in the ordinary course of business, they reflect the influence of all of the incentives and institutions that occur in that practice setting, and so are free from any distortion that might be introduced by researcher intervention.

Researchers who analyze prestructured or hand-collected archives must overcome two challenges. First, the records in the archive were not created with their particular research

⁶ For recent reviews of this literature, see Beyer, Cohen, Lys and Walther [2010] and Dechow, Ge and Schrand [2010].
question in mind, so it can be hard to devise measures that precisely capture the constructs specified in the theory they seek to test. Similarly, prestructured archives reflect many structuring decisions about what data to include and how to aggregate it, so valuable information in the original records may have been lost. Some of this information can be recovered through hand-collection, but the challenge remains. Second, the lack of intervention by the researcher can make it difficult to attribute observed associations to specified causal relationships among theoretical constructs. We now provide examples of studies that make effective use of four opportunities to address these challenges that are available to archival researchers: hand collection, transformation, sample selection and econometrics. We summarize our analysis and recommendations in Table 4.

5.1.1 Hand Collection. Hand-collecting an archive is a particularly direct way of tailoring measures to a research question. General-purpose archives may be structured in such a way that data is aggregated to a level that excludes the information of interest to the researcher. Also, the researcher’s construct might be so specific that no data provider has chosen to include it in a general-purpose archive. In that case, the researcher can hand-collect archival data and structure it for their own purposes.

Larcker and Zakolyukina (2012) illustrate the power of structuring hand-collected data to examine a specific causal construct. The authors seek to test whether managers speak differently in conference calls when managers are being deceptive. Because there wasn’t a preexisting structured archive measuring speech behavior in conference calls, Larcker and Zakolyukina obtain an archive of conference call transcripts and structure it themselves, using algorithms to count word choices and other speech patterns that prior literature suggests are markers of deception. Larcker and Zakolyukina also need a measure of deceptiveness in financial reporting. They begin with a structured archive that lists the dates of each firm’s accounting restatement announcements and the quarters affected, and then use that information to extract and interpret text from hand-collected archives of 8-Ks, AAERs, and other disclosures to create measures of restatement severity. With this customized data, Larcker and Zakolyukina support their predictions that deceptive CEOs and CFOs use more references to general knowledge, fewer non-extreme positive emotion words, and fewer references to shareholder value.

Building hand collected archives can be labor-intensive and expensive. One way to reduce that cost is to rely on another researcher’s archive. For example, Lucian Bebchuk shares
structured data sets originally created for a variety of his published studies on corporate governance issues. Researchers can use this data to examine the causes and consequences of corporate governance across a wide range of firms, allowing generalizability but sacrificing the customization that comes with hand-collecting their own archive.

Another way to deal with the cost of hand collection is to trade statistical power and generality for rich contextualization. As an example, Schwartz-Ziv and Weisbach (2013) acquired the detailed board meeting minutes of eleven firms in which the Israeli government had a large financial interest. They code these meeting minutes to measure the different types of interaction that arise in the boardroom, and offer a detailed look into the behavior of officers and executives. According to their analysis of transcripts, boards spend the vast majority of the time monitoring management, but rarely ever disagree with the CEO. This deep dive into text documents provides a wealth of context that would be excluded from structured data and thus not available otherwise. However, the small and potentially unusual sample of firms raises questions about how well results would generalize across firms operating in other jurisdictions and under more independence from government.

5.1.2 Transformation. Once a researcher has a structured database (whether or not they hand-collected some of it themselves), they almost always find it helpful to create better proxies for their causal constructs by mathematically transforming existing variables into new ones. For example, a researcher who seeks to measure the immediate market response to a disclosure can sum daily market returns over the three days after the disclosure. Such simple transformation is possible when the data is more granular than the construct of interest.

Transformation also can be used to distill measures of the constructs of interest from measures that individually capture very different constructs. For example, starting with Jones [1991], studies in the earnings-management literature apply assumptions about earnings processes, reporting behaviors, and accounting institutions to convert the statistical behavior of CRSP, Compustat and I/B/E/S variables into measures of earnings quality. The tradeoff is that greater transformation typically requires stronger assumptions. For example, Dechow and Dichev (2002) define a broad measure of (low) accrual quality by measuring accruals that are unexplained after accounting for past, current and future cash flows. This allows them to impose

---

8 Transformations are also useful in making data conform more closely to the assumptions of statistical tests (e.g., the log transformation for skewed data), but statistical analysis is not the focus of this paper.
relatively weak assumptions about statistical patterns in earnings and cash flows, but they cannot
distinguish low quality arising from the firm’s underlying volatility from that arising from poor
reporting practices or intentional manipulation. Gerakos and Kovrijnikh (2013) impose stronger
assumptions to uncover intentional manipulation (e.g., making assumptions about managerial
intent). They build their measure by assuming that unmanipulated earnings follow a first-order
autoregressive process, meaning that shocks to earnings persist over the long run and are not
associated from one period to another. They also assume that firms manage earnings to conceal
volatility. Given these assumptions, they are able to interpret associations between earnings
shocks two periods apart as evidence of earnings management. This transformational technique
allows the authors to measure earnings management for a broad cross-section of firms, allowing
strong power to detect associations and many opportunities to demonstrate generality. However,
any evidence that a transformed measure of opportunistic reporting causes or is affected by
another construct becomes a joint test of the assumptions underlying the transformation and the
causal hypothesis.

5.1.3 Sample Selection. The second challenge of relying on archival data (structured
previously or hand-collected) is inferring a causal relation from association-based data without
otherwise intervening in the setting. One useful technique for shoring up attribution in archival
studies is to select sample data that is likely to be particularly informative. The “difference-in-
differences” design is among the most popular sample selection approaches. In this design,
archival researchers restrict their sample to a time just before and after a key independent
variable changes for reasons that are unlikely to be caused by the dependent variable. The
sample also includes observations with varying levels of a key moderating variable. Theory
predicts that difference in the dependent variable will differ as a function of the moderating
variable. This design allows researchers to approximate “the experimental ideal” (Angrist and
Prischke, 2009) of a pretest-posttest design while using the moderating variable to control
effectively for omitted variables that are likely to affect all observations equally.

For example, Balakrishnan et al (forthcoming) wish to test whether a lack of other
information channels causes firms to issue more guidance about their future earnings. To ensure
that the information environment causes guidance choices, rather than the other way around, they
focus on the periods immediately before and after the closure of 43 brokerage houses. Because
the closed houses each covered a representative and arguably random set of firms, and the
closures were closely tied to broader economic conditions, closure is very unlikely to have been effected by managers’ guidance choices. The authors show that the closures were followed by a bigger increase in guidance among firms whose coverage was affected than among firms whose coverage was not affected (a difference-in-differences effect), providing strong evidence for the specified theory, and particularly for the direction of causation.

Researchers can further strengthen causal attribution by narrowing the sample of observations they include in a difference-in-differences design. Iliev (2010) uses a “regression discontinuity” approach to identify the causal effects of Sarbanes-Oxley’s Section 404 provisions on internal controls. Iliev not only restricts his sample to the time immediately before and after Sarbanes-Oxley’s Section 404 provisions came into effect, but also selects only firms whose public equity float is very close to the $75 million cutoff that determines the date they needed to comply with Section 404. He finds that firms just above the cutoff exhibited better earnings quality, at the cost of much higher audit fees. Iliev can attribute these effects quite convincingly to Sarbanes-Oxley because there is little reason to expect differences between firms with such similar float, other than that they are on different sides of the regulatory cutoff.

Careful selection of samples can also help archival researchers focus on the influence of the causal forces of interest in their theory while reducing the influence of other variables. For example, Granja (2014) seeks to test theories of how public disclosure and periodic on-site inspection affect banking failures, and of how economic conditions affect public support for such public disclosures. To focus on these constructs, and reduce the impact of the many other forces that affect banks and voters (without having to gather and include a host of control variables), Granja focuses on a very particular sample: US banks operating in the late 19th and early 20th century, and votes on banking regulation in Indiana and Illinois in 1898. While the sample was selected in part due to the presence of strong exogenous shocks, the simpler nature of banking, regulation and politics in that time period also allows the author to avoid concerns about omitted variables. The banking regulations passed in the period under study affected only a few aspects of banking, while modern regulations include numerous provisions. The votes in 1898 were similarly narrowly focused. Like most studies that select narrow samples, Granja’s provides a wealth of rich contextual information that can help readers interpret the results and develop new theory.
Sample selection is a powerful tool for enhancing attribution and providing rich context. However, researchers cannot always find a sample that includes data immediately before and after an exogenous shock, or observations immediately above and below a key threshold. Even if they can, narrow samples can make it hard to generalize results to the wide range of settings and observations that are excluded from analysis. Sample selection techniques are most helpful when available narrow samples are sufficiently well-suited to the research question, and the sample is sufficiently representative, that the benefits of attribution outweigh the cost of decreased generalization.

5.1.4 Econometrics. Finally, researchers can strengthen attribution in archival studies by using econometric methods to control for omitted variables. The tradeoffs associated with this approach are familiar – these methods typically require the researcher to impose assumptions on the nature of the data and the behavior being examined, and sacrifice context, to obtain the stronger associations and clearer generalizations that come from using large data sets that cover broader and more representative circumstances. Sufficiently sophisticated econometrics can allow strong attributions by demonstrating that the underlying theory can predict a rich structure of causal associations that would be highly unlikely to be observed if the theory were not true. For example, Ecker, Francis, Olssen and Schipper (2014) seek to examine how aggressive earnings management affects shareholder returns and CEO compensation. Ecker et al. clarify the nature and direction of causation by specifying a detailed model linking a wide variety of mediating and moderating variables which they create through a variety of statistical transformations. By showing predicted associations not only between the cause and effect of interest, but also among various mediating and moderating variables, such structural equations reduce concerns about reverse causality and alternative mechanisms. Also, results that exploit the benefits of econometric methods use widely-available measures on a comprehensive sample of firms, and so may be more general than those based on exogenous shocks arising in carefully-selected samples. However, sophisticated econometrics rely on a number of statistical assumptions, so (as with the transformation of variables) most conclusions rely on a test of the joint hypotheses that those assumptions and the causal theory are both true.
5.2 Field Studies

Field studies are similar to archival studies in that researchers don’t intervene in the setting they are examining: they allow dependent variables, variation in independent variables, and the setting in which behavior occurs to arise in the ordinary course of business. Unlike archival studies, researchers do not delegate either of the two distillation tasks when they conduct a field study: they record their observations and structure those records themselves in light of the theory they have specified. Therefore, power as well as attribution can be enhanced by the researcher’s ability to customize variables to precisely capture the constructs of interest. Also, field studies often yield rich contextual insights that help researchers interpret results and guide specification of new theory. However, because researchers do not intervene in the setting, field studies present many of the same attributional challenges as archival work. And, as discussed previously, data could be distorted by researcher involvement in data recording if practitioners in the field setting act differently because they know they are being observed. In field studies, data recording is typically limited by resources and the need for permission from cooperating organizations. As a result, field studies often have relatively small sample sizes and so may lack power for tests of association.9

The definition of field studies derived from our framework is narrow. Many studies commonly called field studies are classified by the framework as analyzing prestructured or hand-collected archives, because researchers have delegated data recording to the organization that was willing to share their proprietary but pre-existing data archives. Many others do not fall within the scope of our paper, because they are devoted to theory building rather than theory testing.10 However, financial reporting researchers often engage in field work, which involves many of the same considerations as a field study. Therefore, we start with a discussion of field work, and then go on to discuss field studies and the choices researchers can make when applying that method.

5.2.1 Fieldwork. As described by the organization theorist John Van Maanen, “in its broadest, most conventional sense, fieldwork demands the full-time involvement of a researcher over a lengthy period of time…and consists mostly of ongoing interaction with the human targets

---

9 For additional discussion of field methods, see Bruns and Kaplan (1987).
10 Many of these studies are published in journals not included in Table 2, such as Accounting, Organizations and Society.
of study on their home ground” (Van Maanen [2011, p. 2]). The researcher “lives with and lives like the subjects being studied” (ibid, p. 2). Such ethnographic fieldwork typically is qualitative and relies heavily on rich detailed description by the researcher. Intensive fieldwork is widely undertaken in anthropological and sociological research that embeds researchers within defined environments that range from tribal cultures in the Amazonian rain forest (Chagnon [1984]), to Japanese fish markets (Bestor [2004]), to Wall Street investment banks (Ho [2009]).

Accounting researchers typically don’t embed themselves in organizations, but they still may follow the spirit of ethnographic work to guide their research. They can visit the premises of the organizations they hope to study, meet with key players to conduct formal interviews and engage in casual conversation, and read the documents those players rely on for decision-making. This fieldwork helps the researcher specify theories that capture the most important institutional factors that shape the behavior to be examined, provide rich context for the results to be reported, and determine what phenomena in the environment are worth recording. Fieldwork also facilitates access to proprietary data by enabling researchers to build the relationships, legitimacy and trust necessary for practitioners to provide access.

Field-based interaction with practitioners is a critical underpinning of many studies, even when data gathering itself occurs via another method. For example, Groysberg, Healy, and Maber [2011] first conduct fieldwork by examining analyst evaluation reports used at an investment bank to help them identify performance metrics used in practice. They then base their primary analyses on a combination of a proprietary hand-collected archive of compensation data obtained from the bank and a variety of publically available archival data to determine key drivers of analyst compensation. They find that analyst compensation is driven more by recognition by clients and the Wall Street Journal than by forecast accuracy. It is likely that the recognition component of an analyst’s production function had largely been overlooked by researchers because of a tendency to focus on the production elements, like forecast accuracy, that exist as publically available archival data. By employing fieldwork and acquiring a

---

11 Our focus is on fieldwork used to facilitate theory testing. Another important role of fieldwork is to describe practice settings from the perspective of the people who participate in them, rather than superimposing a researcher’s views. As Geertz [1973] notes, ‘What we call our data are really our own constructions of other people’s constructions of what they and their compatriots are up to’. Understanding those perspectives is a primary goal of interpretive (or qualitative) research (Ahrens and Chapman [2006]). Other studies use fieldwork in the service of historical analyses, which Carr (1961) defines as “a continuous process of interaction between the historian and his facts.” Such studies can be extremely valuable additions to accounting literature, but lie outside the scope of this paper, because they do not seek to test theory by attributing statistical associations to causal constructs.
proprietary set of compensation records, Groysberg et al. were better able to leverage preexisting archival data and ascertain analysts’ actual incentives.

5.2.2 Field studies. Compared to hand-collected archival studies, the distinguishing feature of field studies is that the researcher records their own observations. Undertaking this task allows researchers to link their measures very closely to the key constructs of the theory being tested, but because the task is laborious, it is often hard to draw enough observations from enough different settings to establishing convincing associations and generalizable results. Table 5 summarizes our analysis and recommendations.

Soltes (2014a) demonstrates the value of collecting records to capture theoretical constructs. He conducts a field study to test a variety of theories about when analysts meet privately with managers of a large publicly-traded firm. Pre-existing archival data about such meetings did not exist, so Soltes persuaded a firm to record information about meetings according to his specifications as meetings occurred over the course of a year. The records include the date and time of each meeting, as well as the analyst’s name and employer. Soltes combines these newly-gathered records with existing structured archives to demonstrate a variety of associations predicted by existing theory: e.g., firm representatives are more likely to meet with analysts who have less experience, produce more frequent forecasts, and work for banks with an underwriting or other close relationship with the firm. Such associations would be difficult to identify without a researcher tailoring the key records to their theory. For example, data obtained by surveying executives to elicit their recollections of meetings would be unlikely to be sufficiently reliable, given the frequency of meetings and the specificity required to test the theory (including the date and time). Likewise, clever transformations of existing structured databases might provide indirect measures of analyst-firm relationships, but (as discussed in section 4.1) such transformations would require the researcher to impose additional assumptions about analysts, firms and the data, and likely lack the specificity provided by Soltes’ direct observations.

The benefits of customized data records inherent in field studies come at a price – it is costly to obtain a sufficient number of independent participants and settings to record, which can

---

12 Note that undertaking the task of recording does not require authors to write or type the records themselves. Authors can direct the tasks conducted by a research assistant (or in this case, an executive assistant employed by the firm being studied), or even direct the production of algorithms used to generate records automatically. The essential requirement is that the data is recorded prospectively in light of the researcher’s theory.
reduce power and makes it difficult to generalize results. Field studies typically require the cooperation of an organization that permits the researcher access to its premises, people and records. Because consent is hard to obtain, researchers must choose their cooperating organization carefully to ensure that they can gather sufficient data to provide strong tests of association and generalizable results. Sample sizes in field studies are often small, but it doesn’t necessarily take that many records to allow for sufficiently powerful tests of association. For example, Soltes’s records capture 75 private meetings over the course of a year. Placed in the context of the 27 analysts covering the firm over the sample period, and considering that the statistical analysis draws inferences from the meetings that did not occur, this sample is large enough to allow associations to be identified with reasonable power.

Generalizability requires establishing that similar results would be obtained in other settings. If Soltes had recorded thousands of meetings with hundreds of analysts, his tests of associations would be more powerful, but would still reflect analyst interactions with only a single firm. The results would be more easily generalized if records of meetings with analysts were gathered from dozens of different firms, but this would require an impractical level of fieldwork. Instead, Soltes makes two design choices that increase generalizability without additional fieldwork.

First, Soltes increases generalizability by focusing his recording efforts on a typical member of the population (the firm). By choosing a large and mature firm publicly traded for over 25 years on the NYSE, Soltes makes it more likely that analysts’ interactions with it are representative of how they interact with other firms.

Second, Soltes concentrated his recording efforts on the setting that would allow greater independence among the decision-makers he wishes to study. We characterize this design choice as gathering data from a station through which many independent decision-makers travel, rather than gathering data from a decision-maker that travels through many stations. Consider the alternative approach of collaborating with an equity research house to gather records of their analysts’ meetings with the firms they cover. This design would raise more questions about generalizability, because the research house may have a distinctive style, perhaps even enforced by company policy. Analysts from other houses might behave differently. Under the view that meetings are instigated by analysts, rather than by the firms the analysts cover, Soltes’ design increases the independence of the decision-makers—a range of analysts seek meetings shortly
after disclosures and shortly before they publish their reports. Moreover, recording events at a station, rather than recording the actions of travelers, addresses a concern about field-based methods raised by Heckman (1992), and discussed by Banerjee and Duflo (2009): organizations willing to provide consent for recording proprietary information may differ from those that do not. Such a concern is particularly salient in research on financial reporting, because the same forces that influence a firm’s willingness to be transparent to stakeholders are likely to influence their willingness to be transparent to researchers. If that firm is providing access as a station, rather than as a decision-maker, this concern is somewhat alleviated.

Few researchers conduct fieldwork that involves “living with and living like” the decision-makers they wish to understand, and even fewer use that fieldwork to conduct field studies as we define them, recording data in ways tailored to their theories. We think this state of affairs is unfortunate, and hope it changes. However, we know that accounting researchers employing all methods engage in lighter forms of fieldwork, such as talking with practitioners and regulators, reading source business and legal documents, trade periodicals and the financial press, and engaging with organizations similar to the target settings they seek to study. Such activities help researchers gather data more intelligently and design more meaningful tests of theory. They also provide useful guidance for readers who wish to place the results of those tests in context, and who wish to conduct follow-up research. That fieldwork can best guide readers and future researchers if it is reported as part of the study. More consistent and systematic reporting of fieldwork will also allow accounting researchers to establish norms for the extent and nature of fieldwork that should underlie empirical exercises.

5.3 Field Experiments

Field experiments are similar to field studies, except that the researcher manipulates independent variables within the field setting. Field experiments share with field studies the benefits of rich contextualization, because all data-generating phenomena arise in the ordinary course of business within a practice setting. The researcher sacrifices some of these benefits by intervening in the setting to manipulate independent variables, but these sacrifices are typically small, because manipulation creates levels of independent variables that deviate only slightly

---

13 For a discussion of using field work in the context of prestructured archival research, see Soltes (2014b).
14 For helpful suggestions about methodology for field experiments see List (2011).
from situations that would have arisen without intervention, and also because the researcher observes responses that occur naturally, rather than eliciting a dependent variable. In exchange for that small sacrifice, the researcher strengthens causal attribution tremendously by randomly assigning decision-makers to the conditions created by manipulation of independent variables. Table 6 summarizes our analysis and recommendations.

Field experiments share most of the challenges of field studies, but the intervention inherent in field experiments makes it harder to gain consent from cooperating organizations because of concern that the organization or its stakeholders could be inconvenienced or harmed in the course of the study. Few firms will consent to manipulations that may harm stakeholders, no matter how small and distantly-related to the firm, or even that create the appearance of harm, unless it involves taking actions the organization otherwise would have taken in the ordinary course of business. Collaborating closely with organizations in the field therefore makes it unlikely that the researcher will cause harm, but likely restricts the topics the researcher can investigate and the manipulations the research can employ. On the other hand, many field experiments rely on little or no consent by any organization, other than their own employers’ approval. These studies are easier to conduct, but the researcher cannot rely on natural checks and balances to keep them from causing harm.

5.3.1 Field Experiments Requiring Sponsor Cooperation. As an example of a field experiment involving high cooperation with a sponsoring organization, Duflo, Greenstone, Pande and Ryan (2013) conduct a field experiment that involves very close collaboration with the Gujarat Pollution Control Board (GPCB) in India to test the influence of auditor independence on pollution control. Prior to the researchers’ interventions, firms hired their own pollution-output auditors, and those auditors operated with relatively little oversight. With the cooperation of GPCB, the researchers assigned each of the 473 polluting firms in two populous areas of Gujarat to one of two settings. Control firms operated exactly as before. Treatment firms were assigned an auditor who was paid by the GPCB. The researchers also introduced random testing by an independent technical organization, as a basis for assessing audit quality. Because they randomly assigned firms to treatment and control, they can be confident in

---

15 Studies that supplement naturally-occurring response variables with elicitations are effectively adding a survey to their field experiment. Studies that look only at elicited responses would not meet our definition of a field study, but our analysis and recommendations would be similar to those provided in our discussion of elicited dependent variables in sections 5.4-5.6.
attributing the improved audits of the treatment group to greater auditor independence. However, the research required extensive fieldwork, not only to devise powerful intervention that would mesh with preexisting institutions but also to secure the cooperation of GPCB, which would presumably be reluctant to take any action that would harm any of the many parties with an interest in the audit process, including firms, auditors, citizens, consumers, politicians, and the GPCB itself.

It can be difficult for authors like Duflo, et al to gain the consent of organizations like the GPCB for their studies. Financial reporting researchers are likely to face even greater challenges, for several reasons. As Duflo (2006) notes, field experiments in developing countries are less costly and more practical than those in developed countries, partly because institutions themselves are less developed and interconnected. Financial reporting involves some of the most developed and interconnected institutions in the world, which raises the cost of interventions that assign people or firms to different conditions, increases the chances that even a simple manipulation will have unintended consequences, and makes it hard to gain consent from those with the decision-making authority. Cooperation likely is easier to obtain when, as in Duflo et al., the experimental design tests the effectiveness of an improvement compared to the status quo. An organization likely would be less receptive to sponsoring a field experiment that seeks to justify the status quo by comparing its effectiveness to a treatment that is hypothesized to be inferior.

5.3.2 Field Experiments Not Requiring Sponsor Cooperation. As an example of a field experiment involving no cooperation with a sponsoring organization, Heinrichs (2014) conducts a field experiment in which she examines which type of financial market participant is more likely to gain private access to management. Heinrichs sends two identical e-mails to over 2,500 firms seeking additional information about the firm and requesting a time to speak, and manipulates between emails whether the sender is a “consultant” or an “investor.” Heinrichs finds that a majority of firms are willing to speak, but access is lower for the “consultant” requesting a meeting than for the “investor.” By manipulating e-mails between firms, Heinrichs is able to offer strong causal evidence that it is the identity of the sender that affects the firms’ willingness to speak, and by observing who responds across a large cross section of firms, she offers powerful evidence of association that is highly generalizable.
Because she does not obtain informed consent from her participants (which would likely distort responses enough to severely hinder attribution), Heinrichs must take great care to minimize any harm her interventions might cause to respondents or other investors. In making this determination, she can rely on her Institutional Review Board and the many similar studies that have been conducted to address similar questions, such as how different names and faces affect responses to applications for employment or college admission (e.g., Bertrand and Mullainathan 2004), and argue that the importance of the insights provided by her research is sufficient to justify the minimal inconvenience she causes for the firms she emails. Researchers must be particularly sensitive to the potential that harm may come not only to the participants in experiments, but others who are indirectly affected. For example, Heinrichs’ study could slow participants’ responses to sincere inquiries. While this effect will be negligible in this case, other experiments might have more substantial impact.

5.4 SURVEYS

Surveys require the researcher to record and structure data elicited from respondents in writing or via interviews. Researchers only intervene by eliciting dependent variables and other information – the researcher does not manipulate independent variables or control the data-collection setting in other ways.

Because survey respondents typically provide data about their naturally-arising practice setting, surveys offer a great opportunity for contextualization, generating rich descriptive data about practitioners’ beliefs and preferences and illuminating previously unhypothesized facts and associations that offer opportunity for theory building. Surveys also can be used to create or augment archival databases with information that isn’t available by other means, using responses to measure independent and dependent variables that then are analyzed using conventional statistical methods. Table 7 summarizes our analysis and recommendations.

One challenge of survey research is that limited intervention makes it difficult to attribute associations to the constructs of interest and the specified causal relationships between them. Biased data could result from researcher choice of samples as well as respondent self-selection to participate. Also, because the researcher does not randomly assign respondents to different levels of manipulated independent variables, self-selection biases might produce significant

---

16 For a helpful discussion of survey methodology in financial reporting, see Nelson and Skinner [2013].
associations due to reverse causality, and because the researcher does not control the setting, theory testing may be vulnerable to omitted variables that are correlated with the measures of interest. Other challenges arise because survey data can be biased by how questions are asked, by respondents’ lack of self-insight concerning their own judgment processes, and by respondents’ desire for particular conclusions. These potential sources of distortion are reduced when respondents report concrete facts, but researchers need to be aware of them when interpreting results.

Graham, Harvey and Rajgopal’s [2005] survey of more than 400 CFOs provides an example of the interesting descriptive insights a survey can provide. Graham et al. provide evidence that CFOs focus on accounting earnings more than cash flows, are willing to sacrifice economic value to smooth earnings or hit earnings targets, and take these actions more to influence share price and further their own career than as a response to debt covenants, credit ratings, political concerns or employee bonuses. These beliefs and preferences are difficult to infer from archival data, so eliciting CFOs’ testimonials about their behavior is valuable. Graham et al.’s survey is largely descriptive, but they map results into the existing literature to help resolve disagreements within it. Given a relatively low response rate and the potential for various other response biases to affect results, Graham et al.’s most compelling insights aren’t with regards to the specific levels of the variables they observe (e.g., that 86.3% of respondents believe meeting benchmarks builds credibility with the capital market), but instead emerge from relative comparisons (e.g., that the dominant reasons to meet or beat earnings benchmarks relate to stock prices).

Baber, Gore, Rich and Zhang [2013] is an example of researchers making effective use of preexisting survey data to augment archival data for theory testing. They use data from a 2001 governance survey by the International City/County Management Association as well as from the Census Bureau’s Annual Survey of Governments. Baber et al. provide evidence that municipal debt costs are higher following disclosure of a financial restatement, and that those costs are mitigated by strong audit oversight and by various provisions that encourage voter participation in governance of municipal financial reporting. Given that sophisticated survey respondents are a scarce resource, it makes good sense for Baber et al. to leverage existing survey data. However, the tradeoff is that they cannot customize the survey questions in light of
their own theory; so, as with other archival research, the quality of attribution depends on the pre-existing structure and content of the data.

Indjejikian and Matejka [2009] provide an example of using a survey to test a theory by gathering specific facts that aren’t available otherwise. The authors first develop a theoretical model of optimal CEO and CFO compensation-contract structure, and then test the implications of that model using data about CFO compensation practices that they obtain from a survey of AICPA corporate members. Results provide evidence that, relative to private entities over the same time period, public entities reduced the proportion of CFO bonuses that were contingent on accounting-based financial performance, which Indjejikian and Matejka’s model suggests occurs as a response to increased costs of financial misrepresentations. Surveys don’t manipulate independent variables, so Indjejikian and Matejka take steps similar to those used in archival studies to strengthen causal attribution from their data, such as using a difference-in-differences design to preclude alternative explanations for results.

Nelson, Elliott and Tarpley [2002] provide an example of a survey that uses relatively open-ended questions to build a database. They survey 253 auditors to elicit 515 experiences those auditors had with clients who attempted to manage earnings. Nelson et al. code indicator variables for each experience to reflect the precision of the financial accounting standards relevant to the transaction, whether the client engaged in transaction structuring (altering the timing, amount or structure of transactions), and whether the auditor thwarted the client’s attempt, and then test hypotheses about determinants of management attempts and auditor adjustment decisions. The open-ended nature of some questions allowed rich contextualization, but requires additional data structuring by the authors, similar to researchers who deal with hand-collected archives. Nelson et al. employ independent coders to reduce the potential that researcher biases could affect this data structuring. Also, given that these data are based on auditors’ self-reported experiences, attribution may be distorted by selection or self-presentation biases, so Nelson et al. are careful to consider potential auditor self-presentation biases when interpreting their data.

Sometimes researchers augment survey data with a small number of follow-up interviews that shed additional light on results. For example, Graham et al.’s [2005] Appendix A includes anecdotes from CFOs describing operational decisions designed to manage earnings. That supplemental information provides valuable contextualization, but researchers must be careful to
acknowledge the small sample from which it is drawn and fight the urge to “cherry pick” only those anecdotes that best support the researcher’s conclusions.

5.5 Laboratory Studies

Laboratory studies are defined by two forms of intervention. Like surveys, laboratory studies allow the researcher to elicit a dependent variable that is customized in light of their theory, but don’t involve the manipulation of independent variables. In addition, laboratory studies allow the researcher to control the setting in which decisions take place. That control allows the researcher to limit the influence of variables that are not specified as causal forces in the theory being tested, while highlighting the influence of those that are.¹⁷

Laboratory studies present two challenges to researchers: the lack of manipulation makes causal attribution difficult, and the controlled setting limits the ability to provide rich context. Thus, researchers considering a laboratory study must think carefully about two decisions: whether their theory can best be tested without manipulating independent variables, and if so, how tightly to control the laboratory setting from which data are collected. Table 8 summarizes our analysis and recommendations.

5.5.1 When to conduct a laboratory study. Laboratory studies are most likely to be an appropriate choice when manipulation of independent variables is difficult and when researchers are seeking to understand broad-ranging psychological or economic forces that do not rely too heavily on participants’ expertise or experience. If manipulation is easy, the researcher would almost surely be better off conducting a laboratory experiment because of the attributional benefits that method enjoys. Also, if participants’ expertise or experience plays a central role in the theory being tested, the researcher typically will need to use a case that is sufficiently similar to practice settings to invoke that knowledge, so the experimental setting will need to be less abstracted from the practice setting to which the theory applies. However, the loss of control that occurs when studies use high-fidelity materials is quite costly to the laboratory study, which cannot rely on manipulations to rule out the effects of omitted variables.

One of the most famous laboratory studies in economics meets these conditions quite well: the seminal paper on speculative bubbles by Smith, Suchanek and Williams [1988], hereafter SSW). SSW seek to understand the conditions that might give rise to speculative

bubbles. It was not feasible for SSW to manipulate variables, because economic theory provided little guidance at the time. Existing theories were quite helpful in specifying how competitive trading could eliminate bubbles, but far less helpful in specifying how competitive trading could generate them. Explanations typically involved a complex mix of behaviors that are inconsistent over time and expectations that are inconsistent with outcomes.

Rather than specify a particular (and surely flawed) theory to test with a manipulation, SSW construct a setting that provides clear measures of bubbles and the behaviors and expectations that drive them, and then analyze the results of numerous renditions of that setting. In each rendition, participants trade an asset that pays a random dividend with expected value D after each of N trading periods, so that its fundamental value declines linearly from DN in period 1 to D in period N. Trade takes place in a double-auction market, allowing the researchers to observe not only the market-clearing prices, but also the bids and offers that generate those prices. By structuring the setting to include an unambiguous fundamental value, SSW could provide an unambiguous measure of price errors. Once they observed a rendition in which bubbles reliably formed and popped, they elicited traders’ expectations of future prices to clarify the causal mechanism. The study revealed that expectations adapt slowly to past information; as participants gain experience, their expectations are more accurate and bubbles are diminished. This study has been followed by a number of laboratory experiments that used SSW’s results to guide choices of independent variables to manipulate.

Another reason why a researcher might choose to avoid manipulating independent variables is when manipulation is likely to distort participants’ behavior due to demand effects or other reactions to researcher intervention. For example, Hobson, Mayew and Venkatachalam [2012] conduct a laboratory study to test a psychological theory predicting that misreporting generates cognitive dissonance that is detectable through measurements of vocal stress. Their study requires an elicitation of both misreporting and oral responses that can be coded to measure vocal stress. While it would be valuable to manipulate whether or not participants report honestly or deceptively, such a manipulation is notoriously difficult to implement without affecting the cognitive dissonance that arises from actually misbehaving (as opposed to bluffing in poker, which is viewed as part of the game). Instead, the authors allow misreporting to arise as a natural choice of each respondent, and use their careful elicitation of dependent variables, and the carefully controlled setting, to demonstrate that misreporters exhibit detectable voice stress.
Hobson et al. highlight one of their key causal constructs in their laboratory setting by asking participants to recall as many of the 10 Commandments as possible, an activity previously shown to increase attention to personal moral codes, and amplify the influence of the resulting cognitive dissonance among those who choose to misreport.

Laboratory studies provide a valuable complement to archival work because their control of the setting improves causal attribution but lacks the easy generalization to practice settings that archival studies enjoy. Partly for this reason, Hobson, et al. pair their laboratory study with an archival analysis of voice stress exhibited by executives in conference calls that occur in the ordinary course of business. The results of the laboratory study strengthen the authors’ attribution of archival associations to their theory, while the similarity of the results when triangulated across the two methods strengthens the authors’ claim that the laboratory results generalize to their target setting.

5.5.2 How tightly to control the setting. SSW control their setting by limiting participants’ opportunity to employ their knowledge of market behavior outside the laboratory setting, and limiting the value of such knowledge. Such controls allow economic forces to dominate decision-making [Smith, 1982], by eliminating the effect of theoretically irrelevant “baggage.” Baggage is particularly harmful to laboratory studies because those studies lack manipulations that can be used to eliminate the effects of omitted variables. SSW limit the chances that participants carry baggage into the laboratory setting by using neutral and largely content-free labels to describe the setting. They limit the value of that baggage to participants by basing all incentives on outcomes that occur within the laboratory setting. For example, the only benefit to purchasing a share is that it might be sold later at a higher price or might pay a higher dividend. We call such settings “self-contained”.

The alternative to a self-contained setting is a scenario setting, in which the researcher describes a hypothetical circumstance that might occur in the outside world, and asks the participant how they would behave. Scenario studies impose less control over the setting, but allow the researcher to activate participants’ knowledge and “peer into the minds of experts” (Libby et al. [2002]) to better understand how experienced practitioners behave in practice settings. Thus, scenario-based laboratory studies are appropriate when researchers want participants to bring some types of “baggage” with them but also want to control the setting in other ways to enable clearer attribution.
As an example, consider the following question from Graham, et al [2005] cited in the previous section: “7. Hypothetical scenario: Near the end of the quarter, it looks like your company might come in below the desired earnings target. Within what is permitted by GAAP, which of the following choices might your company make?” The question is followed by a variety of possible choices, like “decrease discretionary spending” and “repurchase common shares.” While Graham et al. is normally considered a survey paper, this particular question meets our definition of a scenario-based laboratory study, because it elicits a dependent variable from a setting that is controlled in a way that highlights the forces the authors wish to examine while limiting the influence of other factors. By asking participants to assume that they have a desired earnings target, and that unadjusted earnings fall just below it, the researchers are placing the respondent in a (metaphorical) laboratory that controls the setting to some extent. Yet, participants’ responses still can be affected by their beliefs about the nature of their firm and its investor clientele, the options it has available to manage earnings, and many other factors that are not controlled by the researchers. Graham et al. are willing to sacrifice some control to invoke participants’ real world knowledge and examine its effects, even though doing so reduces the researchers’ ability to rule out the effect of nuisance variables.

5.6 LABORATORY EXPERIMENTS

Laboratory experiments require the researcher to perform all of the tasks shown in Table 3. The researcher constructs a setting in which independent variables are manipulated, dependent variables are elicited, and other unstructured data is recorded and structured in whatever manner the researcher believes will enable the best test of hypotheses. Performing all of these tasks enables lab experiments to make particularly strong attributions about causal relations within a theory, because manipulated independent variables preclude reverse causality and elicited dependent variables allow tests of theoretical constructs what wouldn’t be observable otherwise. The tradeoff is reduced contextualization, because the abstraction and control which provide strong causal attributions in lab experiments necessarily reduce the value

---

18 We use the term “lab experiment” to refer to experiments in which participants provide data within a constructed setting (e.g., operationalizing one or more scenarios in which participants make judgments). Lab experiments do not have to occur within a laboratory, but do differ from field experiments that utilize natural settings in which participants may not know that they are providing data for a study. For recent helpful discussions of lab experimental methodology in financial reporting, see, e.g., Bonner [2007], Elliott, Hodge, Kennedy and Pronk [2007], Kachelmeier and King [2002], and Libby, Bloomfield and Nelson [2002].
of descriptive evidence, tests of economic significance, and exploratory analyses. Thus, lab experiments are very well suited for testing the hypothesis that a change in X produces a directional change in Y, but are less well suited for describing the levels of X and Y that arise naturally or for exploring how the change in X might also affect another construct, Z, that the experiment was not designed to investigate.

While lab experiments differ on many dimensions, we single out two that are particularly important methodologically: (1) whether the experiment requires the use of sophisticated participants, and (2) whether the experiment examines individual or aggregate behavior. Table 9 summarizes our analysis and recommendations.

5.6.1 Sophistication of Participants. Like laboratory studies that “peer into the minds of experts”, some experiments require sophisticated participants because their objective is to test a theory that involves significant knowledge of accounting, incentives or institutions that participants have acquired in practice settings. For that knowledge to be applicable, the researcher must construct a setting that captures the key features that require its use. For example, Rose, Rose, Norman and Mazza [2014] examine whether independent corporate directors’ decisions are affected by friendship ties between the director and the CEO that either are or are not disclosed. Given Rose et al.’s focus on how directors’ decisions are affected by the balance of their real-world incentives, Rose et al. obtain access to experienced corporate directors for participation in the experiment. Rose et al. construct a setting in which directors must assess how large of a cut in R&D spending they would approve, with larger cuts having negative consequences for the company but increasing the CEO’s bonus. Between participants Rose et al. manipulate whether directors are told they have (1) no friendship ties with the CEO, (2) friendship ties that are not disclosed, and (3) friendship ties that are disclosed. Consistent with prior psychology research, results reveal that friendship ties increase CEO-friendly decisions, especially when those ties are disclosed, even though directors incorrectly believe disclosure increases their objectivity.

The fact that Rose et al. manipulate friendship ties and disclosure rather than setting those variables at a single level is what makes Rose et al. a laboratory experiment rather than a laboratory study. Their constructed setting includes a relatively high-fidelity case and dependent variable to invoke directors’ real-world knowledge and minimize noise, enabling detection of significant associations with a relatively small sample size. Their clean manipulation of
friendship ties provides more precise attribution by excluding other factors potentially captured by the broad proxies used in prior archival research (e.g., labeling as friends those directors who live in the same state or who are members of the same social club or board). Rose et al. appropriately do not attempt to interpret the levels of R&D cuts as indicating the levels that would occur in practice, given the abstraction necessary to operationalize the experiment, but they do triangulate their results with prior archival research examining R&D cuts and friendship ties.

Other lab experiments address important financial reporting questions that don’t require the use of such sophisticated participants, because they address hypotheses that depend on more basic knowledge. These experiments can utilize more available participants and more abstract settings that facilitate powerful tests of association and strong attribution. For example, Hodder, Hopkins and Wood (2008) test predictions about how the structure of the statement of cash flows affects investors’ cash flow forecasts. Hodder et al. predict that the "reverse order" used by the indirect-method statement of cash flows in practice (NI - Δaccruals = CFO) makes it harder for investors to learn the time-series properties of operating cash flows and accruals, particularly when CFO and Δaccruals are of opposite signs. Testing this hypothesis requires only that participants understand the relation between cash flow, accruals and net income, so Hodder et al. can use 2nd-year MBA students in their experiment. Participants use a simplified statement of cash flows to forecast cash flows for 16 simulated companies, with half of the participants assigned to a reverse-order format and half to a more intuitive "forward-order" format (CFO + Δaccruals = NI). Consistent with expectations, forecast error and dispersion are higher under the reverse format, and opposite signs appear more problematic for participants.

Hodder et al. can construct a setting that includes relatively high abstraction because they are not attempting to invoke complex real-world knowledge of sophisticated practitioners. This approach provides strong attribution, with experimental manipulations producing clear evidence of causality, and other factors that could affect forecasts either controlled, omitted, or spread across treatment conditions via random assignment. Although the “forward-order” condition doesn’t exist in practice, it provides a benchmark against which to judge the effects of the reverse order, as well as providing evidence about an alternative format that standard setters might consider.
Of course, Hodder et al.’s use of abstraction and control necessarily reduces their ability to contextualize their results. A natural question is whether a similar pattern of results would be observed for real-world analysts who learn cash-flow relationships and forecast cash flows in more complex practice settings. Typically experiments triangulate with other papers to provide convergence with other methods or rely on demonstrations of successful applications to provide that evidence, but Hodder et al. provide convergent evidence within the same paper. Specifically, they analyze I/B/E/S data to provide evidence that analysts have higher forecast error for mixed-sign than same-sign cash-flow information, both cross-sectionally when controlling for firm differences and longitudinally when using the firm as its own control. While Hodder et al. note that this archival analysis is vulnerable to various potential correlated omitted variables, the fact that it converges with the results of their highly controlled experiment enhances readers’ faith in the overall set of results provided by both studies.

Experiments that test fundamental psychological processes can inform our understanding of financial reporting without requiring participants to possess any specific accounting knowledge. For example, Rennekamp (2012) examines how non-professional investors respond to readability of non-technical disclosures, so she can use participants obtained from the general population and use a relatively abstract dependent variable. Rennekamp manipulates whether a press release conveys good or bad news and whether it is written and presented in a manner that is relatively more or less readable, varying several dimensions of readability simultaneously to obtain a strong manipulation. She finds that, controlling for comprehension of the disclosure, participants respond more to the sign of the news contained in more readable disclosures and view managers issuing those disclosures as more credible, which triangulates with archival research (e.g., Miller [2010]) indicating that disclosure readability affects the behavior of small investors.

5.6.2 Individual v. Aggregate Behavior. The lab experiments discussed above all draw predictions from psychological theory. However, accounting also has a rich tradition of lab experimentation based on economic theory. These studies typically test predictions that depend on participants’ incentives and the institutional arrangement in which they interact, so they tend to require that the researcher construct tightly controlled settings that abstract unnecessary information to focus on economic behavior. These studies typically focus more on participants’
actions rather than their beliefs, particularly when they are focusing on aggregate dependent measures, such as share prices and trading volumes, that are determined by those actions.

Qu [2013] illustrates how a focus on aggregate behavior can guide design choices. Qu tests the effect of communication on a coordination problem similar to that facing multiple investors who each have the option to contribute additional funding to a bank, but the contribution pays off only if enough other investors contribute as well. Coordination failures arise when the bank is strong enough to survive but fails because each investor worries that the others won’t contribute. Qu manipulates whether investors communicate indirectly through market prices or directly through unverifiable “cheap-talk” disclosures about their intentions. She finds that market disclosures do not reduce coordination failures, because the same uncertainties that create those failures also lead to low prices for the bank, which serve as self-fulfilling prophecies. Cheap-talk disclosure is effective, however, because it costs investors nothing to disclose optimistic intentions, and their incentive is to live up to their intentions when enough other investors have expressed similar intentions.

Qu’s theory is drawn from a branch of economics (coordination games) that has a strong psychological component due to its emphasis on how players form expectations of other players’ actions. However, Qu’s focus is not on the underlying psychological processes, which might be better studied by presenting scenarios to one participant at a time. Instead, Qu focuses on how coordination outcomes are affected by the disclosure of prices or reports. Because the mechanism of the theory operates through disclosure, she elicits actions (trading, reporting and investment) and does not focus on beliefs or judgments. Because the theory is general enough to apply to any group of people operating within the institutions she creates, she need not use expert bankers and investors. Rather, she can train participants to be experts within the entirely self-contained setting she creates. This “local expertise” allows her participants to learn from experience how their actions affect outcomes and payoffs in that setting, and helps ensure that results are not influenced by extraneous factors, such as people’s beliefs about the nature of banking risk in the global economy. Such carefully constructed settings allow strong attribution, but of course trade off contextualization given the many ways in which they abstract from practice settings.
6. Building a Literature with Multiple Methods

As discussed in section 5 and summarized in Tables 4-9, every method provides advantages for researchers to exploit and challenges for researchers to overcome. The ultimate contribution of achieving each goal cannot be determined in isolation, however, because it depends on the state of the existing literature. For example, a literature might provide extensive evidence of associations between variables, but little sound basis for causal attribution. In that case, a laboratory experiment can contribute by allowing clear attribution to the key constructs said to drive the associations arising in the practice setting, even if it uses a highly controlled setting that limits contextualization and generalization. Similarly, a field study might contribute useful context to interpret previously observed associations, even if its small sample size limits its power to test its own reported associations.

To illustrate how theories and methods combine within an important literature in financial reporting, we focus on the large body of research that examines the consequences and effects of recognizing information in financial statements rather than disclosing it in a footnote (i.e., the “recognition v. disclosure” literature). We draw insights and specific studies partly from excellent reviews of this literature by Schipper [2007] and Libby and Emett [2014]. We then discuss how researchers can enhance complementarity by reporting the results of their studies more completely.

Early studies on recognition and disclosure used laboratory experiments to show that liabilities have less effect on individual judgments when disclosed than when recognized. Harper et al. [1987] show this effect for pension liabilities among both bankers and students. Belkaouei [1980] demonstrates a similar effect for bankers and accountants when assessing environmental liabilities, as does Wilkins and Zimmer [1983] for lease liabilities. Subsequent experiments generalize these findings to non-liabilities and to other differences in information location (e.g., Hirst and Hopkins [1998], Maines and McDaniel [2000], and subsequent studies provide evidence that information about unrealized investment gains and losses has less of an effect when it is recognized on a non-performance statement than when it is recognized on a performance statement). By manipulating whether information is recognized or disclosed (or whether it is recognized in less accessible statements), these studies allow us to attribute differences in judgments to information location. By examining the judgments of various groups
of professionals across a variety of settings, the studies taken together suggest that their results are generalizable across laboratory settings and to professionals in practice settings.

The laboratory experiments on individual judgments did not establish that results would generalize to aggregate market outcomes, which are disciplined by strong competitive forces (Gonedes and Dopuch [1974]). One way to provide that evidence is by examining whether individual firms appear to make reporting decisions that anticipate that recognition vs. disclosure matters. For example, Lys and Vincent [1995] use hand-collected archival data to analyze a single event, AT&T’s acquisition of NCR. Lys and Vincent conclude that AT&T paid a large premium to report higher income to investors, even though this income effect is easily backed out by examining contemporaneously available disclosures. While Lys and Vincent did not create their own records, their paper largely plays the role of a field study in the literature on recognition and disclosure by providing rich contextual information about the motivations of the actors and the responses of investors, and makes it more plausible that laboratory evidence for the effect of recognition generalizes to a market setting—or, at least, that managers would expect it to.

Lys and Vincent provide rich contextualization by sacrificing generality: they provide evidence on only a single firm. A number of subsequent archival studies demonstrate differential effects of recognition and disclosure in highly competitive U.S. equity markets. For example, Davis-Friday et al. [1999, 2004] provide archival evidence that OPEB liabilities receive lower valuation weights when disclosed than when recognized. Ahmed et al. [2006] provide similar evidence for fair values of derivatives. By using large archives of publicly-available data, these studies provide powerful associations directly in the target setting. While the lack of researcher intervention reduces these studies’ ability to make causal attributions, the successful triangulation between laboratory and archival data allows convincing attribution to the causal effect of recognition and disclosure from the literature as a whole.

Additional research has focused on clarifying the mechanisms by which recognition causes stronger responses than disclosure. One line of research focuses on the psychological mechanisms underlying individual responses. For example, Clor-Proell and Maines [2014] argue that disclosed numbers are less reliable, and so should be weighted less, and support this theory by providing laboratory experimental evidence that public (but not private) company managers expend more effort estimating recognized v. disclosed liabilities. Hirst et al. [2004] argue that,
controlling for reliability differences, disclosed numbers are less accessible, and so are weighted less regardless of how they should be weighted. Their laboratory experiment provides evidence that bank analysts are better able to recognize interest rate risk and value differences when fair values are recognized in the balance sheet and income statement than when partly disclosed, even though the analysts indicated no difference in reliability. These studies employ sophisticated participants who complete tasks that include key features of their practice settings but that control, abstract or manipulate others, allowing strong attribution of observed effects to theory.

Field experiments outside accounting providing converging evidence. For example, Brown, Hossain and Morgan (2010) conduct field experiments using popular online auction platforms in Taiwan and Ireland. Brown et al. sold identical iPods, holding constant that shipping charges were disclosed in the detailed description of the item but varying the amount of shipping charges and whether the charges were also indicated in the title of the item listing. Brown et al.’s results indicate that sellers can earn more revenue by charging higher shipping charges that are less prominently disclosed. Their field experimental evidence enables strong attribution of observed behavior to their disclosure manipulation, and the fact that their results are observed for people spending significant dollars in a familiar practice setting helps to generalize and contextualize results produced in laboratory experiments.

A second line of research considering causal mechanisms examines the institutional forces that allow aggregate market outcomes to reflect biased individual responses, rather than eliminating them through the discipline of competition and arbitrage. Bloomfield and Libby [1996] use an experiment to show that prices in laboratory markets respond more strongly to information that is more widely available. While this self-contained study is free of virtually all context, its strong controls, manipulations and elicitations clearly establish the mechanisms of market discipline. This study, along with others like it, provide the foundation for a relaxation of the Efficient Market Hypothesis (Fama, 1970), called the Incomplete Revelation Hypothesis (Bloomfield, 2002). A large body of subsequent archival studies have relied on and supported its prediction that information most costly to process is less completely revealed in market prices, a mechanism consistent with the stronger market effects of relatively accessible recognized information.
A third line of research focuses on managers’ motivations for choosing accounting methods that recognize good news and disclose bad news. Dichev et al. [2013] provide survey evidence that a broad range of CFOs view net income as critical, implying that they think recognition and disclosure decisions matter to investors and affect aggregate market outcomes. Hodder and Hopkins [2014] analyze a hand-collected archive of comment letters sent to the FASB, and conclude that bank managers expect markets to respond more strongly to bad news about loan values if it is recognized rather than disclosed.

In total, these papers produce compelling evidence of significant associations that can be cleanly attributed to theory and that generalize to multiple settings as well as bridging individual-level behavior and market-level outcomes. No one paper or method makes all of these contributions, but each makes complementary contributions that combine to allow this literature to provide important insights and develop over time.

7. Conclusion

We hope this paper will encourage researchers to gather new data, and will guide them to do so more efficiently and effectively. Those who gather new data are building lampposts that allow them to find insights they think will best contribute to the literature, rather than looking under existing lampposts to find whatever contributions are already illuminated. Future researchers may also be able to look under these new lampposts, particularly when data gathering results in data sharing.

We also hope that this paper will encourage people to focus less on research methods and more on how researchers contribute effectively to a literature by identifying research goals that will advance the literature, choosing data gathering tasks to achieve those goals and making wise decisions in light of the tasks they undertake or delegate. We have spent many hours mapping published literature into our framework, and have found that it is rare that a single study, much less an author, can be accurately described by one of the methods so commonly referred to (e.g., “a survey” or “a survey researcher”). Moreover, all of us have found that the framework has changed how we analyze and evaluate research. We are now less likely to require that all papers achieve all goals, given the tradeoffs inherent in methodological choices. We are now more likely to focus on the particular goals a paper is emphasizing and how those goals mesh with a researcher’s methodological choices to contribute to the extant literature. This focus has helped
us appreciate the complementary nature of different approaches to data gathering, and makes us want to reward researchers for undertaking “tedious and difficult” data gathering that pushes the literature forward. We will view this paper as a success if it helps readers do the same.
References


Figure 1
Example of a Theory Specified for Empirical Testing
Figure 2
Libby-Box Framework for Depicting How Empirical Studies Test Theory
Figure 3
How Do Decisions About Data-Gathering Tasks Accumulate to Define Methods?

This figure depicts how seven research methods can be defined according to the choices researchers make with respect to five data gathering tasks. Squares indicate decisions of whether to undertake a data gathering task (yes) or delegate it to others or allow it to occur without researcher intervention (no). Circles indicate methods. Thus, the “prestructured archival” method is defined by choosing “no” for all five data gathering tasks; other methods are defined by choosing “yes” for one or more tasks. “IV” refers to independent variables; “DV” to dependent variables. See Table 3 for definitions of data gathering tasks and methods.
Table 1
Goals of Empirical Theory-Testing Literatures

An empirical theory-testing literature is successful if it can explain the causes and consequences of a wide variety of phenomena in settings of interest. This table identifies five goals (broken into subgoals) that must be achieved to attain such success.

1) Specification: To specify theoretical constructs, and how and why they are causally related.
   a. Specify at least two constructs that may be causally related. Researchers may also wish to extend and refine theories, as indicated in Figure 1, by adding distal causes or effects, adding moderating or mediating variables, or revise existing constructs.
   b. Specify causal relationship between constructs (direction and possibly magnitude)
   c. Specify mechanism driving causal relationship (mediation and/or moderation)

2) Association: To test whether statistical associations between measured variables are consistent with theoretical predictions.
   a. Maximize Power (to detect associations when they exist)
   b. Minimize Bias (to assess sign and magnitude of associations accurately)
   c. Test Statistical Significance (to determine whether sign of the association is driven by chance)
   d. Verify Analysis (to determine that the tests are conducted as reported)

3) Attribution: To attribute statistical associations to the specified causal theory.
   a. Clarify Direction of Causation (to rule out that the cause is influenced by the effect)
   b. Narrow Definition of Causal Construct (to discriminate between similar constructs)
   c. Show Evidence of Mechanism (to show that the causal connection is driven by the reasons specified by theory)
   d. Account for Peripheral Variables (to show that associations are driven by the constructs attributed as causal factors)

4) Generalization: To indicate that causal attributions will generalize to other circumstances.
   a. Show Robustness (reproducing similar results using new techniques with the same data)
   b. Show Replication (reproducing similar results using similar techniques with new data)
   c. Show Triangulation (reproducing similar results using different methods on new data)
   d. Show Application (reproducing similar results applying the causal theory in practice settings of interest)
5) **Contextualization:** To provide context that clarifies interpretation of the results and guides future research.
   
   a. Report Descriptive Statistics (to clarify the distributions of variables being examined)
   
   b. Report Economic Significance (to clarify the impact of the effects on matters of interest)
   
   c. Report Unhypothesized Associations (to guide future research seeking to attribute unhypothesized associations to new theories)
   
   d. Report Unhypothesized Facts (to guide future research seeking to understand the contexts and behaviors in settings of interest)
Table 2
Research Published in JAR, JAE & TAR, 2003-2013

This table reports the numbers of paper published in the *Journal of Accounting Research*, the *Journal of Accounting and Economics*, and *The Accounting Review*, by area (i.e. paper type) and the type of data used in the analysis. The data type classification is based on the primary data utilized in the analysis that served to support the hypotheses in the paper.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Financial Reporting</th>
<th>Audit</th>
<th>Managerial</th>
<th>Tax</th>
<th>Governance</th>
<th>Methodology</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival</td>
<td>607</td>
<td>88</td>
<td>38</td>
<td>70</td>
<td>113</td>
<td>25</td>
<td>941</td>
</tr>
<tr>
<td>Survey &amp; Case-Based</td>
<td>7</td>
<td>4</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>“Lab” Experiment Data</td>
<td>53</td>
<td>39</td>
<td>39</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>137</td>
</tr>
<tr>
<td>Field Experiment</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Analytical &amp; Simulation</td>
<td>79</td>
<td>13</td>
<td>48</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>165</td>
</tr>
<tr>
<td>Total</td>
<td>748</td>
<td>144</td>
<td>136</td>
<td>87</td>
<td>123</td>
<td>33</td>
<td>1,271</td>
</tr>
</tbody>
</table>
Table 3
Methods and the Tasks the Researcher Undertakes

This table defines seven common empirical methods according to the choices that researchers make with respect to five data gathering tasks. For each task, researchers can choose to undertake the task themselves (yes) or delegate the task to others or allow it to occur without researcher intervention (no). Record data indicates whether researchers acquire and record data observations for themselves, or delegate that task to others. Structure data indicates whether researchers, after acquiring data, organize the data into tabular form prior to empirical analysis, or delegate that task to others. Elicit dependent variable indicates whether researchers acquire data for dependent variables by posing questions, or passively observe data. Manipulate independent variables indicates whether researchers randomly assign participants to alternative treatments, or allow independent variables to vary naturally. Control the data-generating setting indicates whether researchers use abstraction and control to affect the data-generating process, or allow that setting to arise naturally. Section 3.1 describes each of these tasks in more detail. Section 3.2 defines the different empirical methods in more detail.

<table>
<thead>
<tr>
<th>Does the researcher…</th>
<th>Pre-structured Archival Analyses</th>
<th>Hand-collected Archival Analyses</th>
<th>Field Studies</th>
<th>Field Experiments</th>
<th>Surveys</th>
<th>Laboratory Studies</th>
<th>Laboratory Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record data?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Structure data?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Elicit dependent variables?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Manipulate independent variables?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Control the data-generating setting?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 4
Prestructured and Hand-collected Archival Analyses: Analysis and Recommendations

**Definition:** Researcher delegates recording and possibly structuring of data archives that have been drawn from a practice setting without intervention.

**Advantages to exploit:** Large samples allow high power to test for associations. Intervention-free data from practice setting strengthens generalization and contextualization.

**Challenges to overcome:** Attribution is difficult without customized distillation or intervention.

**Recommendations:**

To test more specific constructs, transform variables when overly strong assumptions are not required; structure data in light of theory when hand-collected archives are available.

To establish causal mechanisms, select samples when exogenous shocks serve as natural manipulation in a sufficiently representative setting to support generalization; use structural equations with large samples and clear theory.

To control for nuisance variables, select narrow samples when sufficiently representative, or use econometric techniques with large samples and clear theory.
Table 5
Field Studies: Analysis and Recommendations

**Definition:** Researcher records and structures data observed in a practice setting without intervention.

**Advantages to exploit:** Ability to customize recording allow measurement of specific constructs. Intervention-free data from practice setting strengthens generalization and contextualization.

**Challenges to overcome:** Small samples limit power to test for associations. Small number of settings studies limits generalizability. Attribution is difficult without intervention.

**Recommendations:**

To strengthen attribution and contextualization, conduct initial fieldwork to guide subsequent data recording and structuring.

To strengthen generalization, gather data from stations through which many decision-makers pass, rather than decision-makers who pass through many stations.
Table 6
Field Experiments: Analysis and Recommendations

Definition: Researcher records and structures data observed in a practice setting after manipulating independent variables.

Advantages to exploit: Manipulation allows strong causal attribution. Ability to customize recording allow measurement of specific constructs. Limited intervention in practice setting strengthens generalization and contextualization.

Challenges to overcome: Small samples limit power to test for associations. Small number of settings studies limits generalizability. Attribution is difficult without intervention. Researcher must avoid harming participants or others because of effects of real-world intervention.

Recommendations:

To strengthen attribution and contextualization, conduct initial fieldwork to guide subsequent data recording and structuring.

To strengthen generalization, gather data from stations through which many decision-makers pass, rather than decision-makers who pass through many stations.

To avoid harming participants or others, obtain consent from cooperating organization that can be trusted to protect stakeholder interests, or exercise great sensitivity to those affected by the burden of data gathering and the impact of manipulations.
Table 7
Surveys: Analysis and Recommendations

**Definition:** Researcher records and structures data elicited from participants without other intervention.

**Advantages to exploit:** Ability to elicit dependent variables and other information allows measurement of specific constructs. Limited intervention strengthens generalization and contextualization.

**Challenges to overcome:** Lack of manipulated dependent variables makes causal attribution difficult. Participant responses may be inaccurate or dishonest.

**Recommendations:**

- To strengthen attribution, focus on variation in responses across subsamples, and tailor questions to capture specific constructs.

- To strengthen contextualization and limit biases that reduce the validity of associations, be careful about how questions are asked and use independent coding for qualitative questions.
Table 8
Laboratory Studies: Analysis and Recommendations

Definition: Researcher records and structures data elicited from participants within controlled setting.

Advantages to exploit: Ability to elicit dependent variables and control setting allows researcher to capture specific constructs.

Challenges to overcome: Lack of manipulation hinders attribution. Controlled setting limits contextualization.

Recommendations:

Verify that manipulation isn’t possible, due to lack of theory or impracticality.

Create self-contained setting to study the effects and mechanisms by which individual decisions affect aggregate outcomes; use realistic scenario settings to evoke practitioners’ knowledge and beliefs.
Table 9
Laboratory Experiments: Analysis and Recommendations

**Definition:** Researcher records and structures data elicited from participants in response to manipulating independent variables within controlled setting.

**Advantages to exploit:** Ability to elicit dependent variables and control setting allows researcher to capture specific constructs. Ability to manipulate variables strengthens attribution.

**Challenges to overcome:** Extensive intervention limits generalization and contextualization.

**Recommendations:**

- Use realistic scenario settings to evoke practitioners’ knowledge and beliefs, use more controlled settings to study psychological and economic forces.
- Create self-contained settings to study aggregation.