Revista de Educação e Pesquisa em Contabilidade
Journal of Education and Research in Accounting

Periódico Trimestral, digital e gratuito publicado pela Academia Brasileira de Ciências Contábeis | Available online at www.repec.org.br

REPeC, Brasília, v. 16, n. 3, art. 5, p. 318-337, Jul./Sep. 2022 | DOI: http://dx.doi.org/10.17524/repec.v16i3.3155 | ISSN 1981-8610

# Quantitative Empirical Research in Management Accounting: A Proposed Typology and Implications for Internal versus External Validity

**Andson Braga de Aguiar**
*http://orcid.org/0000-0003-4034-4134*

**Daniel Magalhães Mucci**
*https://orcid.org/0000-0002-0658-1470*

**Myrna Modolon Lima**
*https://orcid.org/0000-0003-2084-481X*

## Abstract

**Purpose:** The purpose of this study is twofold. First, we propose a typology of quantitative empirical research in management accounting based on two design features: presence of control group and sample representativeness. Second, we discuss implications of the methods for trade-offs between internal and external validity.

**Method:** Based on previous methodological studies we develop a typology with eight quantitative empirical methods.

**Results:** Based on the two design features, we propose eight quantitative empirical methods for management accounting studies: (1) laboratory experiment, (2) crowdsourcing experiment, (3) field experiment, (4) natural experiment, (5) single entity survey, (6) proprietary archival study, (7) large-scale survey and (8) pre-structured archival study. In addition, we critically compare the trade-offs and discuss the implications of these methods for internal and external validity.

**Contributions:** The contribution is twofold. First, the proposed typology can help junior management accounting researchers increase the familiarity with the available empirical methods, some of which are still incipient in Brazil. Second, this study states that the choice of an empirical method typically implies benefits in terms of a validity type (e.g. internal validity) at the expense of other validity type (e.g., external validity). Claims of causality and results generalizability depend on which validity type is prioritized and remedies adopted to increase the overall validity of a study's results.

**Keywords:** Quantitative empirical methods; Management accounting; Control group; Sample representativeness; Validity types.

**Published in Portuguese and English. Original Version in Portuguese.**

Received in 8/16/2022. Accepted on 8/19/2022 by Dr. Gerlando Augusto Sampaio Franco de Lima (Editor). Published on 14/10/2022.
Organization responsible for the journal: Abracicon.

Copyright © 2019 REPEC. All rights, even translation, are reserved. It is allowed to quote part of articles without prior permission if the source is identified.

## 1. Introduction

It is not uncommon for management accounting (MA) researchers in early stages of their careers, and eventually for more experienced researchers, to make claims that cannot be supported by the empirical method used to design the study. For instance, researchers may be tempted to make causal claims in studies that do not meet the necessary conditions for that, such as there is no plausible alternative explanations for the effect other than the cause (Shadish, Cook, & Campbell, 2002). Similarly, researchers may be tempted to make claims of generalizability of results, even though random selection as a major way to provide evidence for generalization is rarely the case (Trochim, Donnelly, & Arora, 2016).

We argue that these flawed claims are made because MA researchers, particularly junior researchers in Brazil, are not familiar with the available empirical methods and, consenquently, with the implications of the different methods in terms of the validity framework. In fact, prior studies have consistently shown that the range of empirical methods used by MA researchers in Brazil is rather narrow, with a prevalence of large-scale surveys (e.g., Aguiar, 2018; Frezatti, Aguiar, Wanderley, & Malagueno, 2015). As long as junior MA researchers in Brazil are not familiar with the available methods, they may not be equally familiar with the relative advantages and disadvantages of each method in terms of research validity.

Moreover, although junior MA researchers can access available empirical methods through methodological books (i.e., Smith, 2022; Saunders, Lewis, & Thornhill, 2019), these books provide only a limited view for at least two reasons. First, research methodology books broadly discuss each available empirical method, such as experiment, archival and survey, not considering that each method can be further dismembered in more specific research designs, with associated validity benefits and threats. For instance, archival studies can use either pre-structured or proprietary data (Moers, 2006). Second, these books usually focus on the protocols to be followed by each broad research method, lacking a more specific and critical discussion about the implications for the validity framework.

The lack of knowledge about the available empirical methods and their implications for research validity may have fundamental consequences for the advancement of a research field. On the one hand, the lack of familiarity with the diverse options of available methods can actually narrow down the scope of research questions that MA researchers investigate as well as overlook the complementarities of using different empirical methods to address the same research question (Bloomfield, Nelson, & Soltes, 2016). On the other hand, the lack of knowledge about the implications of the available methods for the validity framework can lead MA researchers to make unsupported claims, particularly, in terms of causal relationships and generalizability of results.

Given the importance to increase the breadth of empirical methods used by junior MA researchers in Brazil, the purpose of this study is twofold. First, we propose a typology of quantitative empirical research in MA. Quantitative research is prevalent in accounting research (Hesford, Lee, Van der Stede, & Young, 2006; Nascimento, Junqueira, & Martins, 2010; Aguiar, 2018). Moreover, validity criteria for quantitative studies differ in type and importance from qualitative studies. We focus on MA because a typology has already been proposed to help accounting researchers to select appropriate methods based on the data gathering tasks involved (Bloomfield et al., 2016). While useful in general, Bloomfield's et al. (2016) typology does not consider the singularities that MA researchers face in the process of selecting the particular empirical method to be used for the research project. For instance, when selecting a survey, MA researchers have to further decide the level of analysis and representativeness of the sample (Van der Stede, Young, & Chen, 2006). The proposed typology is based on two design features: presence of control group and sample representativeness. Building on these criteria, we discuss about eight different methods: (1) laboratory experiment, (2) crowdsourcing experiment, (3) field experiment, (4) natural experiment, (5) single entity survey (6) proprietary archival study, (7) large-scale survey and (8) pre-structured archival study.

Second, we discuss the main implications of the proposed typology for research design in terms of the validity framework. Validity refers to the "approximate truth of an inference" (Shadish et al., 2002, p. 34). The validity of research findings are inevitably affected by the research design (Dyckman & Zeff, 2014). Each empirical method thus affects differently the types of validity and researchers have to make trade-offs between the validity types. While four validity types exist, we focus on internal and external validity for two reasons. First, internal validity is key for causal claims, while external validity is central for claims of generalizability (Shadish et al., 2002). Second, there generally are trade-offs between internal and external validity that each empirical method has to face (Luft & Shields, 2014; Roe & Just, 2009).

The contribution of this study is twofold. First, we propose a typology of quantitative research in MA based on two key design features that can help expand the toolkit available to junior scholars in the design of their studies. For instance, we highlight that the choice of a survey requires from MA researchers to make a subsequent choice related to the unit of observation between large-scale or single-entity surveys. The empirical methods included in the proposed typology can help junior MA researchers increase the familiarity with the available alternatives, some of which are still incipient in Brazil, such as experimental designs (Aguiar, 2017; Nascimento et al., 2010). Second, the discussion about the implications of the proposed typology provides insights on the trade-offs among the validity types MA researchers have to pay attention when selecting a particular empirical method. For instance, the choice of an empirical method including a control group may favour internal validity while, at the same time, pose challenges in terms of external validity. In other words, the choice of a particular empirical method implies a simultaneous choice of the validity types that is favoured in the study versus creates challenges for the researcher.

## 2. Validity in MA Research and Proposed Typology

### 2.1 Internal Validity and Causal Claims: Control Group

There are three main types of cumulative research studies (Trochim et al., 2016): i) descriptive studies, which focus on documenting what are the key characteristics of a population or phenomenon; ii) relational studies, which focus on examining the relationship between two or more variables; and iii) causal studies, which focus on determining whether one or more independent variables causes one or more dependent variables. The willingness to stablish causal relationships and make causal claims is one of the main goals of quantitative studies in MA (Van der Stede, 2014). In causal studies, the key concern is to increase internal validity, that is, "The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured" (Shadish et al., 2002, p. 38).

For a research project to be able to make causal claims, three conditions have to be met: the cause precede the effect, the cause is related to the effect, and no plausible alternative explanations for the effect are found other than the cause (Shadish et al., 2002; Luft & Shields, 2014). Researchers can use theoretical arguments and/or lagged variables to claim that the cause precede the effect. Researchers can also use different statistical tools to show that the cause is related to the effect. The main challenge with causal claims is though to guarantee the absence of plausible alternative explanations for the presumed causal relationship.

The most effective way to eliminate alternative explanations is by knowing what would have been the effect if the cause had not been present, that is, by creating a counterfactual (Floyd & List, 2016). The creation of a counterfactual is a design choice that involves the presence of a control (or baseline) group. In other words, a **control group** creates a useful counterfactual inference, essential for research studies interested in establishing causal relationships (Lonati, Quiroga, Zehnder, & Antonakis, 2018).

The presence of a control group is the hallmark of experimental studies (Shadish, Cook, & Campbell, 2002; Trochim, Donnelly, & Arora, 2016), so that this design choice allows the separation of experimental from non-experimental studies. As a counterfactual, a control group represents a group of respondents/participants who are comparable to the treatment (experimental) group in every way possible, with the main difference being that the control group is not exposed to the treatment/manipulation (Oehlert, 2003). For instance, in a study examining the effect of rewards on employee motivation, the treatment (control) group would include participants who (do not) receive a reward. For such research design, the researcher would compare employee motivation between the treatment group and the control group used as a counterfactual and examine whether or not the fact that a reward is provided would affect employee motivation.

Overall, the main benefit of including a control group is that researchers can mitigate the likelihood of alternative explanations for a causal relationship and, thus, improve internal validity and more confidently make causal claims (Floyd & List, 2016).

## 2.2 External Validity and Claims of Generalizability: Sample Representativeness

Researchers can gather primary and/or secondary data to address the research question. In either way, researchers have to stablish the process of selecting units from a population of interest, being them individuals, groups, subunits or organizations; that is, researchers have to stablish the sampling process that will allow them to generalize results from the sample to the population from which the units are selected (Speklé & Widener, 2018). Generalizability claims are desirable and common in accounting studies (Dyckman & Zeff, 2014). For that, researchers have to enhance external validity, that is, "the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes" (Shadish et al., 2002, p. 83).

There are two main approaches to select a sample: probabilistic and non-probabilistic. The main difference between the two approaches is that probabilistic sampling involves random sampling, in which every unit in the population has the same chance of getting selected (Dyckman & Zeff, 2014). In order to make claims of generalizability of results, the best approach is the use of probability sampling that, however, is rarely feasible (Speklé & Widener, 2018). Alternatively, researchers can use an approach called proximal similarity model that allows generalizations from the observed sample to other samples based on the degree to which the other samples are similar to the observed sample (Trochim et al., 2016).

Regardless of the sampling approach, the challenge is to obtain **representative samples** so that generalization is possible. Through representative samples, researchers can provide evidence of external validity and then make claims of generalizability of results. Given the challenge with the use of random sampling, MA researchers typically obtain representative samples through the use of heterogeneity sampling, in which units are chosen purposively to reflect diversity on pre-defined important dimensions (Shadish et al., 2002). For instance, in a study examining the effect of reward types on executive myopic behavior, researchers can use pre-structured data from a variety of organizations, that is data that are gathered and stored prior to the beginning of the research (Das, Jain, & Mishra, 2016; Moers, 2006), and are publicly available through platforms such Economatica and Compustat. As another example, a study examining the effects of control systems in family firms on non-family members behaviour could collect representative samples through questionnaires sent to a large number of family firms. Representative samples can also be obtained through the use of proprietary dataset, that is, dataset is confidential and can be accessed only if the data proprietor/owner grants access (Moers, 2006). Examples of proprietary data are third-party surveys (e.g., consulting firms) and firm internal data (Das et al., 2016; Moers, 2006).

Overall, the key benefit of representative samples is improved external validity, which allows for more supported claims of generalizability of results (Trochim et al., 2016).

## 2.3 Proposed Typology

Building on the two criteria—control group and sample representativeness—we propose a typology with eight quantitative empirical method alternatives that MA researchers can use when addressing their research questions (Figure 1). The proposed typology involves two dimensions based on the design choices associated with each criterion.

In the **first dimension**, researchers decide whether the research design includes a control group. If included (Figure 1, left side), the choice leads to the use of experimental empirical methods; conversely, non-experimental empirical methods are chosen (Figure 1, right side). In the **second dimension**, researchers decide whether they will use samples that are more or less representative. For both experimental and non-experimental studies, the researcher can select an empirical method that either has lower or higher sample representativeness. We next discuss about each alternative empirical method.

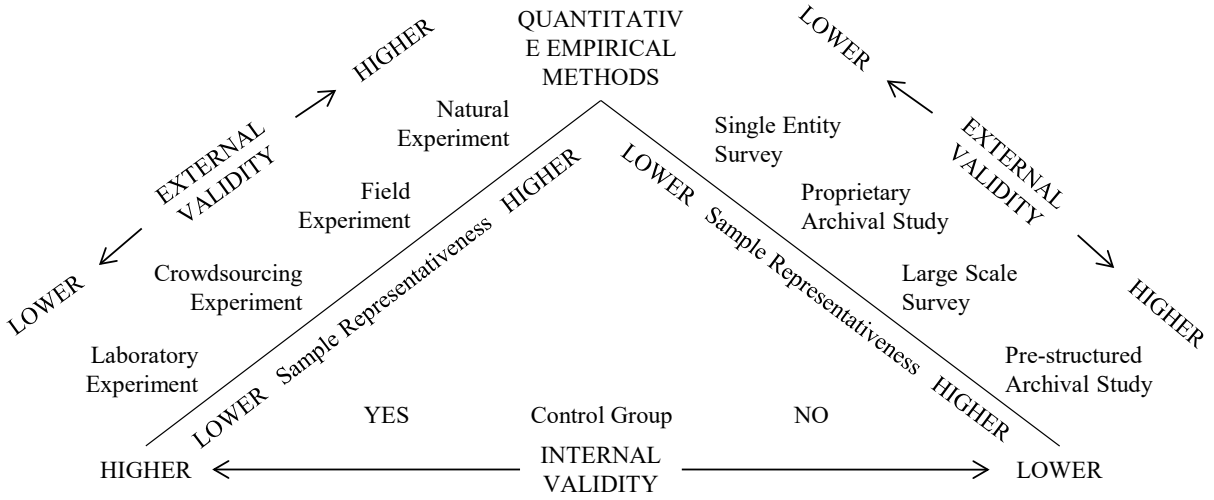


**Figure 1.** Proposed typology of quantitative research in MA

# 3. Alternative Quantitative Empirical Methods

## 3.1 Laboratory experiment (Lab experiment)

Lab experiments have as main characteristics the presence of control group, random assignment, and the fact that the treatment (e.g., independent variable) is manipulated (Bloomfield et al., 2016; Sprinkle & Williamson, 2006; Swieringa & Weick, 1982). Participants in lab experiments typically have similar characteristics—same undergraduate course, approximately same age, and similar background (Shadish et al., 2002; Sprinkle & Williamson, 2006). Lab experiments can be run in labs or, for social scientists, in rooms where the experimenter can ensure physical control over participants, such as rooms with individual spaces that prevent participants from looking to other participants or to use technological gadgets during the experiment.

The fact that lab experiments use control group, random assignment, and manipulated variables improves internal validity for such studies. As already mentioned, control groups create the counterfactual to eliminate alternative explanations (Oehlert, 2003). Random assignment means that two or more groups of units are created that are probabilistically similar to each other on the average, so that any outcome differences that are observed between the experimental and the control groups are likely due to treatment rather than pre-existing differences (Shadish et al., 2002). Manipulated variables enhances that the cause purposefully precedes the effect. By using random assignment and manipulated variables, the researcher can control the research setting and also reduce alternative explanations by isolating the effects of confounding variables (Sprinkle & Williamson, 2006).

The controlled environment of lab experiments, however, contributes to lower external validity of such studies. For commonly recruiting homogeneous participants (e.g., undergraduate students) through non-probabilistic, purposeful sampling (Carpenter, Harrison, & List, 2005), the sample representativeness in lab experiments tend to be quite low. In fact, it is desirable that participants for lab experiments are a *tabula rasa* so that the only influence they experience is the treatment, what makes students well suited, particularly if the study requires specific knowledge (e.g., analysing financial statements) but do not require specific experience (Mortensen, Fisher, & Wines, 2012; Trottier & Gordon, 2018). Accordingly, lab experiments are highly criticized for their artificiality (Harrison, 2005; Carpenter et al., 2005).

An example of a lab experiment that focuses on managers' decision-making process is Haesebrouck's (2021) study on the effects of information acquisition effort and the induced psychological ownership on managers' reporting. As part of their job, managers can either exert great effort to acquire, synthesize and analyse data from several sources or easily acquire this information if the company possess good information sharing systems. To test the predictions, the author used an experimental design, where information acquisition (endowed vs. earned) and saliency of honesty in the reporting context (less vs. more salient honesty) are manipulated. The paper ensures internal validity by following several experimental procedures (i.e., randomization, demographics, and direct observation of participants). The author states that the experiment focuses on theory testing and results might not be generalized to other settings.

Overall, on the one hand, due to the presence of control group, random assignment, and manipulated variables, lab experiments are considered the empirical method with the highest internal validity (Trochim et al., 2016) and with the greatest chance to stablish causal claims. On the other hand, the use of less representative samples and artificial environments makes lab experiments the empirical method with the lowest external validity (Asay, Guggenmos, Kadous, Koonce, & Libby, 2021) and the least chance to make generalizability claims.

## 3.2 Crowdsourcing experiment (or online experiment)

Crowdsourcing experiments are relatively similar to lab experiments due to the fact that they also include control group, random assignment, and manipulated variable. However, the two experimental designs differ as for location and participants. Crowdsourcing experiments use online platforms (e.g., MTurk, Prolific, and CrowdFlower) rather than labs or rooms. The use of such platforms allows researchers to recruit participants from larger pools beyond students from limited geographical locations (Peer, Brandimarte, Samat, & Acquisti, 2017).

As crowdsourcing experiments use control group, random assignment, and manipulated variables, internal validity for such studies is enhanced. However, the use of online platforms creates internal validity threats by decreasing researchers' control over the experiment. The researchers have less control over noises and biases, such as lack of attention or effort, or even unqualified participants (Bentley, 2021). Researchers can also face fraudulent behaviour from participants, affecting data integrity and reliability (Aguinis & Ramani, 2021).

The use of a more heterogeneous group of participants, however, contribute to increase the external validity for crowdsourcing relative to lab experiments due to more representative samples. Online platforms allow researchers to filter several demographic characteristics that can help narrow down participants on the one hand; at the other hand, researchers can examine their theory recruiting participants with different ages, cultures, and backgrounds. Moreover, despite the increasing use of online platforms, there is still great criticism over the reliability of such data compared to other types of experimental data (Chmielewski & Kucker, 2020).

Example of crowdsourcing experiment is Murphy, Wynes, Hahn, and Devine's (2019) study on the internal and external motivation to honest reporting. The authors use MTurk participants in an experiment where they have both opportunity and incentive to misreport in order to test different motivations behind honesty. Since the goal is to test motivation behind decisions, three manipulations are used: baseline (control group) vs. reward vs. punishment. The authors explain that MTurk participants are suitable for the experiments given the nature of tasks (i.e., not specialized). Consistent with crowdsourcing experiments, the study combines internal and external validity. The authors ensure internal validity by applying different manipulations and ensure external validity by using participants with different characteristics.

Overall, on the one hand, researchers give up part of the laboratory experiment's ability to monitor participants in-person, which renders the crowdsourcing experiment lower internal validity compared to lab experiments. On the other hand, the use of online platforms that allow researchers reach a larger and more diversified pool of participants improve the external validity of crowdsourcing experiments relative to lab experiments.

## 3.3 Field Experiment

Field experiments are field studies that use the experimental method in which either treatments or effects are observed for longer periods of time (Lourenço, 2019; Bloomfield et al., 2016). Field experiments are similar to other experiments, including control group, random assignment, and manipulated variable. The main difference is that they are conducted in the field (i.e., organizations) and participants are professionals working in naturally occurring environments, which are generally not aware of being taken part of the experiment (Lourenço, 2019; Floyd & List, 2016). Moreover, field experiments allow researchers to use proprietary data and measured variables to achieve the study's goal (Asay et al., 2021).

Due to the fact that field experiments also use control group, random assignment, and manipulated variables, internal validity for such studies in enhanced. Yet, field experiments face higher internal validity threats due to the lower level of control as researchers cannot create an artificial, isolated environment in which participants will make judgments and/or decisions and variables will be manipulated as can be done in lab experiments (Lourenço, 2019).

The use of naturally occurring environments that are not artificially created by the researcher, in which participants do not know that they are part of the experiment and are performing their daily tasks, contribute to an increase of external validity in field experiments relative to lab and crowdsourcing experiments. Although these features can increase external validity in terms of realism, field experiments are also bounded to less representative samples (e.g., business unit, single company).

Example of field experiment is Cronin, Erkens, Schloetzer, and Tinsley's (2021) study on the effects of controlling failure perceptions on performance. The authors manipulate the video-based message that sales workers see during their weekly meeting in one of 20 Brazilian distributorships of a multinational direct sales organization. In the treatment condition, sales workers see a video message from the regional head encouraging workers to look at failure as a "natural part of history". While, in the control condition, sales workers see the same regional head summarizing the organization's history. The authors conducted the experiment during a four-week period, with control and treatment groups, and with proprietary data (e.g., weekly sales commission). Moreover, even though field experiments have more generalizable results compared to lab and crowdsourcing experiments, the authors disclosed the limitations of using experimental methods and confounding factors associated with field studies.

Overall, field experiments provide a potent combination: benefits of internal validity, consistent with the experimental method, and real world data that comes from the field (Bloomfield et al., 2016; Lourenço, 2019). This combination offers a great mix between "control and realism usually not achieved in the laboratory or with uncontrolled data" (Floyd & List, 2016, p. 438). Field experiments then offer lower level of internal validity due to reduced control associated with less artificial and isolated environments, but these costs come at the benefits of greater generalizability to naturally occurring settings.

## 3.4 Natural Experiment

Natural experiments are naturally-occurring events (e.g., exogenous shocks) that are not manipulated but can establish a "contrast between a treatment and a comparison condition" (Shadish et al., 2002, p. 17). Events that do not happen through natural intervention (e.g., flood, hurricane) can also be treated as natural experiments (e.g., law adoption) (Mcvay, 2011). Thus, natural experiments are feasible when events occur due to human or natural interventions, and researchers can compare ex-ante to ex-post outcomes. While including a control group, natural experiments do not include random assignment and manipulated variables due to the naturally-occurring events. Because of that, natural experiments are regarded as quasi-experiments (Lourenço, 2019; Aguinis & Bradley, 2014). Moreover, natural experiments can benefit from large available pre-structured or proprietary data.

While natural experiments also use control groups as counterfactuals to increase internal validity, this research method tends to face more internal validity threats. The main reason is that control and treatment groups are not designed by researchers, but set exogenously. Then, participants are not randomly assigned to experimental conditions and, consequently, researchers cannot assume that experimental groups are equivalent (Trochim et al., 2016). As such, researchers that run natural experiments have to find alternative methods to minimize internal validity threats by using sophisticated statistics, such as Difference-in-Difference, Regression Discontinuity Design, and Synthetic controls (Lonati et al., 2018).

The extent to which results from natural experiments can be generalized depends on the unit of observation. When the unit of observation is at the organizational level (most common case), that is, when the exogenous event has affected a large sample of organizations, the ability to obtain representative samples and generalize results is higher. Yet, when the unit of observation is at the subunit or individual level, the challenge with external validity is higher, since researchers will have a hard time in convincing that the studied organization is somehow similar to other organizations.

Example of natural experiment in MA is Flammer and Kacperczyk's (2016) study about effects of stakeholder orientation on innovation in company's business decisions. The authors explore the enactment of state-level constituency statutes, particularly the statutory change of the responsibility from shareholders to stakeholders. From 1980 to 2006, 34 US states have adopted constituency statutes that focus on stakeholder value creation. The authors use data about patent creation from the National Bureau of Economic Research (NBER) Patent Data Project from 1976 to 2006 to assess the dependent variable, innovative productivity, measured as the number of patents and citations divided by the number of company's employees. To deal with internal validity threats, the authors collected several control variables that could act as confounding factors. This study is consistent with natural experiments providing higher external validity due to the use of large available pre-structured dataset.

Overall, similar to field experiments, natural experiments combine the benefits of internal validity by including control group, and external validity by examining naturally-occurring events (Bloomfield et al., 2016; Lourenço, 2019). With that, this empirical method provides intermediate degrees of internal and external validity (Roe & Just, 2009) that allow balanced claims of causal relationship as well as generalizability of results.

## 3.5 Single Entity Survey

Single entity surveys are developed through the use of questionnaires and submitted to respondents from one organizational setting. Studies that employ this method do not include a control group, random assignment, or manipulated variables. For single entity surveys, qualitative information from the organizational context (i.e., interviews) has a central role in shaping and "calibrating" research instruments from the survey. Moreover, single entity surveys are commonly administered as a cross-section and the survey instrument can be designed with questions to capture facts rather than perceptions and opinions. This empirical method has better chances of using longitudinal designs as well as better limit target respondents and use random samples. Also, the support of the firm can influence respondents' motivation and minimize potential response and non-response biases. Respondents in single entity surveys typically include employees (individual level), groups (team level), subunit managers (subunit level).

The fact that single entity surveys do not include control group, random assignment, and manipulated variables pose internal validity threats, considering the non-temporal separation of the cause and effect as well as issues related to measurement and survey design such as common method bias (Speklé & Widener, 2018). In particular, self-selection is a relevant internal validity threat, whether in terms of the firm that accepts to participate or the respondents that accept to take part in the survey. The access to the field, however, can reduce the internal validity threats in different ways. For instance, the understanding of the setting before survey implementation allows researchers to obtain more qualified and less biased responses by identifying knowledgeable potential respondents and engaging them to respond the survey. Also, the use of qualitative data can help calibrate the research instrument and allow appropriate choices of ideal employee-titles for the study. If the use of longitudinal designs is feasible, internal validity is substantially improved.

Single entity surveys face challenges in making inferences to other organizations since the research model is context dependent. Although researchers in single entity surveys usually obtain larger response rates than in large-scale surveys (Hiebl & Richter, 2018) and are more capable of addressing issues related to nonresponse bias, the findings are not generalizable to other samples due to the particularities of the context. This means that findings are only empirically applicable to organizations with very similar characteristics and phenomena. Thus, considering these limitations, when discussing external validity for single entity surveys researchers are usually referring to the generalization of the sample to the population of individuals within that organizational setting in certain period of time (e.g., current middle-managers or assembly workers) or the generalization to the theoretical level.

Example of single entity survey in MA is Mucci, Frezatti and Bido (2021), which investigates the association between four enabling design characteristics of budgets and managers usefulness perceptions in an organization that operates in the electric utilities industry. The single entity survey was developed with a sample of 75 middle managers from different business areas (i.e., finance, operations, and marketing) and was operationalized with the support of the firm budgeting manager. The researchers obtained a high response rate of 42% and followed several procedures to mitigate internal validity threats (i.e., strict theoretical model, control variables, and common-method bias).

Overall, single entity surveys face several internal validity threats, but are able to deal with part of them by strictly defining the theoretical and empirical model in light of the context being investigated (Luft & Shields, 2002) as well as partially controlling for confounding effects that emerge from the context. In general, by not having control group single entity surveys present lower internal validity than the experimental methods. Similarly, the limited design in the organizational setting poses external validity threats to single entity surveys. In fact, this empirical method may be the one with the lowest external validity among the non-experimental methods. Then, while single entity surveys have a balanced degree of internal and external validity, the overall degree of validity is lower than the one from field and natural experiments that also have a more balanced degree of validity.

## 3.6 Proprietary Archival Study

Proprietary archival studies use confidential archival data from third-party surveys or firm internal data that can be accessed only if the data proprietor/owner grants access (Moers, 2006). As a non-experimental method, proprietary archival studies do not include control group, random assignment, and manipulated variables. If the dataset is obtained from third-parties, this empirical method can include a large and heterogeneous set of observations that can render greater sample representativeness, which brings proprietary archival studies closer to pre-structured archival study. In turn, if firm internal data is used, this empirical method is closer to single entity surveys due to the potential use of context-specific information that can help researchers in the choice of appropriate proxies for the relevant variables.

As a non-experimental method, proprietary archival studies face several internal validity threats mainly associated with self-selection and endogeneity concerns (Lourenço, 2019). Moreover, the lack of control group makes more challenging ruling out alternative explanations to results. Proprietary archival studies then typically follow econometric procedures to deal with internal validity threats, such as using instrumental variables. These threats are particularly serious when third-party surveys are used. When using firm internal data, researchers can gather additional unstructured data and structure it to create measures that are suitable for addressing relevant research questions to a particular study, which gives more flexibility in the search for suitable proxies for variables of interest. Moreover, proprietary data can be combined with field interviews to help identify suitable proxies for relevant variables.

Proprietary archival studies involve large and comprehensive samples (Moers, 2006). Yet, their ability to make inferences from observed samples to samples located in other places and at other time may be more challenging, particularly relative to firm internal data. When using firm internal data, sample representativeness is a key challenge and thus the ability to generalize results from observed samples to other samples. The reason is that observations at the subunit or individual (e.g., employee) level of analysis may be unique to the sampled firm characteristics, making difficulty to generalize results to firms with different characteristics. Because of that, similar to laboratory experiments, proprietary archival studies can better argue in favour of generalization to the theory being tested.

Ikäheimo, Kallunki, University, and Schiehll (2018) is a proprietary archival study with no control group and using firm internal data. Ikäheimo et al. (2018) examine the relationship between performance-based incentives for white-collar employees and firm future profitability and if this relationship depends on task complexity. They use a large proprietary compensation dataset from a survey administrated by the Confederation of Finnish Industries. The dataset includes over 564,000 individual employee-year and 7,820 firm-year observations over the years 2002–2011. The authors conduct robustness checks to deal with endogeneity issues and thus increase internal validity. Given the large dataset, there is high ability to generalize results to other organizations that use incentive schemes and plan to change them, or do not currently have incentive schemes but plan to adopt them.

Overall, on the one hand, when proprietary archival studies include third-party surveys, internal validity threats are higher due to self-selection bias and endogeneity problems, but the sample representativeness and thus external validity is higher. As we will see, these features bring this empirical method closer to pre-structured archival studies. On the other hand, when internal firm data is used, internal validity concerns are remedied by the use of context-specific information, but external validity is harmed as the dataset is specific for a single organization. In this case, these features bring this empirical method closer to single entity surveys.

## 3.7 Large-Scale survey

Large-scale surveys are commonly used in quantitative MA research (Van der Stede et al., 2006; Speklé & Widener, 2018). This method collects data through the use of questionnaires that are submitted to a broad set of potential respondents; yet, obtaining a large number of responses is usually a challenge for MA researchers. Sample representativeness depends on the use of random selection and the response rate; however, most studies in the MA field use convenience samples. Similar to other non-experimental methods, large-scale surveys do not include control group, random assignment, and manipulated variables. Large-scale surveys are usually designed cross-sectionally and relevant variables are elicited in the research instrument (Bloomfield et al., 2016). Even when involving longitudinal data, large-scale surveys can suffer from the lack of responses since respondents might not be available to participate in second or third survey waves. Finally, large-scale surveys commonly consist of data that express facts, opinions, or perceptions considering different levels of analysis, typically at the organizational or business unit levels.

As a non-experimental method, large-scale surveys face several internal validity threats mostly associated with self-selection bias and endogeneity problems. There is no counterfactual or time difference between cause and effect. When longitudinal data is collected, it is more feasible to argue in favour of time difference between cause and effect, but then mortality is a relevant threat. To deal with internal validity threats, researchers conducting large-scale surveys can adopt several remedies, such as, defining theoretical population, target population and target respondents, using different types of responses (facts, opinions, etc.) and carefully designing the research instrument. Yet, while effective, these remedies cannot completely overcome internal validity threats.

Large-scale surveys benefit from the use of large samples and are thus able to provide evidence that can be generalized to other samples. However, researchers are usually aware that potential for generalizing results depends on sample representativeness, nonresponse bias, and response rates (Hiebl & Richter, 2018; Spekle & Widener, 2018). Also, the ability of large-scale surveys to generalize results depends on whether nonprobability or probability sampling is used, with the former being the most common strategy. Hence, the toolkit available for increasing the arguments for external validity are different to those applied for other methods.

Example of large-scale survey in MA is Bedford, Spekle, and Widener (2022), which conduct a large-scale survey with Business Unit (BU) managers from the Netherlands, with a final sample of 83 respondents. They study how firms change budget tightness in response to global crisis and the implications of budget tightness for employee stress and emotional exhaustion, also considering an enabling budget design as moderator for this relationship. Bedford et al. (2022) run a cross-sectional survey design using online questionnaires, addressed to a population of 172 BU managers. Although sample size is relatively small, they obtain a high response rate of 48.3 percent. The authors follow several procedures to mitigate internal and external validity threats. They address nonresponse bias and common method bias, in addition to use validated instruments and control variables.

Overall, this empirical method faces higher internal validity threats than all previous empirical methods, whether experimental or non-experimental. Compared to single-entity survey and proprietary archival studies, large-scale surveys do not consider contextual data (i.e., organizational particularities), rendering problems related to definition of proper respondents, suitability of research instruments, and biases associated with individual responses (e.g., hallo effect, social desirability, lack of knowledge). On the other hand, compared with the other non-experimental methods discussed so far, results of large-scale surveys have a higher level of external validity due to the participation of a large set of respondents.

## 3.8 Pre-Structured Archival Study

Pre-structured archival studies use archival data that is recorded and structured by third-parties, whose primary purpose is not academic research (Bloomfield et al., 2016; Moers, 2006). Pre-structured archival studies do not include control group, random sampling, or manipulated variables. Relevant variables are operationalized through the use of proxies defined from the available dataset. The dataset typically includes observations for several respondents, mostly at the organizational level. Pre-structured archival studies are the primary research method used in the accounting literature (Bloomfield et al., 2016). For MA studies, the lack of available public data makes more challenging the use of pre-structured archival studies relative to alternative research methods (Hesford et al., 2006; Moers, 2006; Aguiar, 2018).

As a non-experimental method, pre-structured archival studies have difficulty in ruling out alternative explanations as observed associations between variables of interest can be attributed to reverse causality, omitted correlated variables, or miss-specified functional form (Gassen, 2014). Similar to proprietary archival studies, pre-structured archival studies face several internal validity threats associated with selection bias and endogeneity that researchers try to solve by using econometric tools (e.g., instrumental variables) (Lourenço, 2019).

Given that pre-structured archival studies use large samples (Das et al., 2016; Moers, 2006), their ability to make inferences from observed samples to samples located in other places and at other time is relatively high. The unit of observation tends to be highly representative, so that results may be generalizable to the population of interest. However, selection bias may pose challenges for sample representativeness in pre-structured archival studies, in particular, self-selection bias since the disclosure of MA information is not random (Moers, 2006).

Laviers, Sandvik, and Xu, (2021) is an example of pre-structured archival study. It does not include control group and uses several large available datasets. Laviers et al. (2021) examine investor reactions to CEO pay ratio voluntary disclosures. They collect proxy statements from firms listed in the Standard & Poor's 1500 index with mandated CEO pay ratio disclosure and classify firms as having low, middle, or high pay ratio. The authors combine CEO pay ratio information with information collected from database platforms, such as stock returns from CRSP, financial information from Compustat, and executive compensation from Execucomp. Given the lack of control group, Laviers et al. (2021) conduct several additional analysis and robustness tests using different empirical specifications to increase internal validity. Regarding generalizability of results, the authors deal with self-selection by using an estimation procedure by which they include the inverse Mills ratio (Heckman, 1979).

Overall, on the one hand, due to the lack of control group, random assignment, and manipulated variables, pre-structured archival studies face several internal validity threats that are dealt with through the use of econometric tools. In fact, we suggest that pre-structured archival studies are the empirical method with the lowest internal validity. On the other hand, due to the use of large dataset, including observations from organizations at different industries, ages, sizes, and so on, pre-structured archival studies tend to include highly representative samples, increasing the ability of this empirical method to make generalizable claims.

## 3.9 Summary

We summarize in this section the proposed typology of quantitative empirical methods, emphasizing the implications in terms of trade-offs between internal and external validity associated with each method. Figure 2 displays each empirical method positioned according to the relative benefits of internal versus external validity.

It can be observed in Figure 2 that experimental studies have in general higher internal validity due to the presence of control group. In particular, lab experiments are the method with the highest internal validity and the lowest external validity. Moving from lab to natural experiments, there is an increase in external validity due to the use of more representative samples (natural experiments), less artificial environments (field experiments), and more heterogeneous participants (crowdsourcing experiments), associated with a decrease in internal validity caused by the use of less controlled settings (natural experiments), naturally occurring environments (field experiments), and online platforms (crowdsourcing experiments).

It can also be noted that pre-structured archival studies have the highest external validity due to the use of large dataset and the lowest internal validity due to selection and endogeneity problems. We consider that pre-structured archival studies and large-scale surveys have about similar degrees of internal validity, while the external validity is decreasing due to low response rates for large-scale surveys. Proprietary archival studies and single entity surveys are the non-experimental methods with the lowest external validity due to the use of dataset that is specific for a single organization and, at the same time, with the highest internal validity due to the consideration of context-specific information.

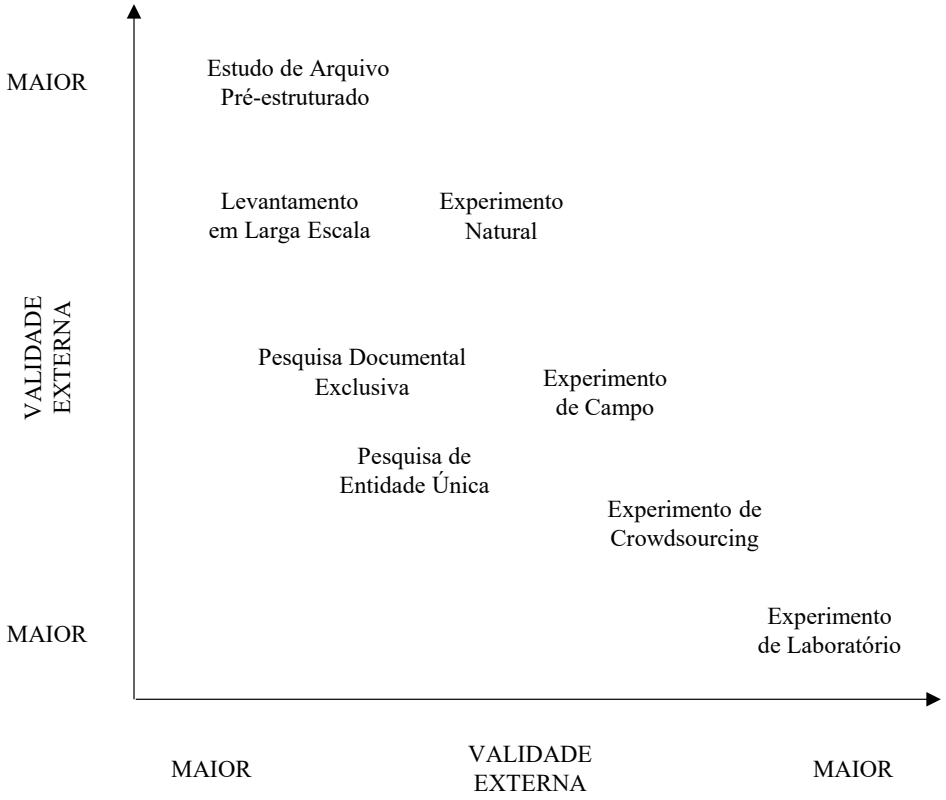


**Figure 2.** Trade-offs between internal and external validity for each empirical method

## 4. Conclusion

This study proposes a typology of quantitative empirical research in MA and discusses the main implications of the proposed typology for the validity framework. In particular, building on two criteria (control group and sample representativeness), we propose eight quantitative empirical methods that MA researchers can use when addressing their research questions, including experimental (lab, crowdsourcing, field, and natural experiments) and non-experimental (single entity surveys, proprietary archival studies, large-scale surveys, and pre-structured archival studies) alternatives. We focus on the implications of the proposed typology for the trade-offs between internal and external validity.

The proposed typology and the validity implications can benefit junior MA researchers in Brazil in several ways. First, this study can help broaden the scope of research questions to be investigated. For instance, if the researcher is interested to examine the role of value statements on employees' behaviour, alternative methods can be used such as designing lab experiments and manipulating value statement to capture employees' responses; accessing single organizations to use available data on employees' understanding of organizational values; or collect survey responses from employees located in different organizations on their perceptions of organizational values and behavioural responses (e.g., goal commitment).

Second, this study can help MA researchers to build research programs involving the sequential use of alternative methods to address the same research question, with the evident benefit of replication and generalization of results. For instance, the behavioural effects of tight budgetary controls examined in large-scale surveys can be further examined in experimental designs to offer stronger evidence of causality. Third, this study can help MA researchers identify alternative methods to examine research questions that are not suitable to be addressed through conventional methods (e.g., large-scale surveys). For instance, if the research question is to examine COVID-19 effects on employees' use of accounting information for decision making, researchers could have access to an organization and collect proprietary data about frequency of use of accounting information prior and post the beginning of the pandemics in a natural experiment and then examine whether this use has been altered.

Finally, we call the attention of junior MA researchers to the implications of the selected empirical method for the trade-offs between internal and external validity. By choosing one method over the other, the researcher is also selecting the relative importance placed on internal versus external validity. For instance, lab experiments increase internal validity, while create challenges for external validity. Conversely, pre-structured archival studies boost external validity, but pose higher internal validity threats. Then, while causal claims are more feasible when lab experiments rather than pre-structured archival studies are used, seeing that the researcher follows appropriate procedures to enhance internal validity, generalizability claims are more feasible when pre-structured archival studies rather than lab experiments are used, again seeing that appropriate procedures to enhance external validity is adopted.

In any case, we recommend that junior MA researchers care about their design choices since "The research findings are inevitably the product of the research design" (Dyckman & Zeff, 2014, p. 697). In particular, two points can help with the choice of the empirical method. First, the choice of the empirical method should not be driven by the availability of a specific dataset or the curiosity in using a specific method but instead by the research question (Kinney 2019; Dyckman & Zeff, 2014). While the exploration of different methods can be valuable for acquiring new research skills, the research question should come in the first place to guide this decision. The benefits in terms of increased chances of publication are higher when researchers obtain deep knowledge on how to use specific empirical methods because each method involves a different protocol to be followed to deal with validity threats. Then, regardless of the method selected, researchers are expected to apply appropriate research protocols, according to the best practices stablished in the area, when conducting the study. Second, researchers will be better off by choosing an empirical method that is feasible, given existing constraints, such as data availability, access to organizations, time, and money.

MA researchers are becoming increasingly creative in how to gather data through the use of mono-method as well as multi-method research designs. For mono-method studies, MA researchers are taking advantage of internal validity associated with experimental studies and external validity associated with archival studies by running quasi-experimental studies, using proprietary archival data and design choices typical of experimental studies, such as pre- and post-measures (e.g., Brüggen, Grabner, & Sedatole, 2021; Forker, Grabner, & Sedatole, 2020). For multi-method studies, MA researchers are combining different data collection procedures (e.g., Bol, Braga de Aguiar, & Lill, 2020; Wouters & Wilderom, 2008). The main benefit of combining different methods is to increase sources of relevant data and provide stronger results by using methods that complement each other, such as proprietary archival data on employee performance combined with perceptual measures on employees' motivation captured through single entity surveys.

While we discuss about the implications of each method for the trade-offs between internal and external validity, we acknowledge that construct and conclusion validity are also critical for establishing valid results. Construct validity refers to "the validity of inferences about the higher order constructs that represent sampling particulars", while conclusion validity refers to "the validity of inferences about the correlation (covariation) between treatment and outcome" (Shadish et al., 2002, p. 38). Overall, these four validity types are interdependent and the use of the predictive validity framework, also known as Libby Boxes, can be helpful for researchers better visualize their research design as well as identify potential validity threats.

## References

Aguiar, A. B. (2017). Pesquisa Experimental Em Contabilidade: Propósito, Desenho E Execução. Advances in Scientific and Applied Accounting, 10(2), 224–244. https://doi.org/10.14392/asaa.2017100206

Aguiar, A. B. (2018). O pequeno mundo da pesquisa em contabilidade gerencial no Brasil: discussão sobre desenhos alternativos de pesquisa. Revista de Contabilidade e Organizações, 12, e151933. https://doi.org/10.11606/issn.1982-6486.rco.2018.151933

Aguinis, H., & Bradley, K. J. (2014). Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. Organizational Research Methods, 17(4), 351–371. https://doi.org/10.1177/1094428114547952

Aguinis, H., & Ramani, R. S. (2021). MTurk Research : Review and Recommendations. Journal of Management, 47(4), 823–837. https://doi.org/10.1177/0149206320969787

Asay, H. S., Guggenmos, R. D., Kadous, K., Koonce, L., & Libby, R. (2021). Theory Testing and Process Evidence in Accounting Experiments. The Accounting Review.

Bedford, D. S., Spekle, R. F., & Widener, S. K. (2022). Accounting , Organizations and Society Budgeting and employee stress in times of crisis : Evidence from the Covid-19 pandemic. Accounting, Organizations and Society, (xxxx). https://doi.org/10.1016/j.aos.2022.101346

Bentley, J. W. (2021). Improving the Statistical Power and Reliability of Research Using Amazon Mechanical Turk. Accounting Horizons, 35(4), 45–62. https://doi.org/10.2308/HORIZONS-18-052

Bloomfield, R., Nelson, M. W., & Soltes, E. (2016). Gathering Data for Archival, Field, Survey, and Experimental Accounting Research. Journal of Accountig Research, 54(2), 341–395. https://doi.org/10.1111/1475-679X.12104

Bol, J. C., Braga de Aguiar, A., & Lill, J. B. (2020). Peer-Level Calibration of Performance Evaluation Ratings : Are There Winners or Losers ?

Brüggen, A., Grabner, I., & Sedatole, K. L. (2021). The Folly of Forecasting: The Effects of a Disaggregated Demand Forecasting System on Forecast Error, Forecast Positive Bias, and Inventory Levels. The Accounting Review, 96(2), 127–152. https://doi.org/10.2308/tar-2018-0559

Carpenter, J. P., Harrison, G. W., & List, J. A. (2005), Field experiments in economics: An introduction. In: Harrison, G. W., Carpenter, J., & List, J. A. (Eds.) *Field Experiments in Economics* (*Research in Experimental Economics, Vol. 10*), Emerald Group Publishing Limited, Bingley, pp. 1-15. https://doi.org/10.1016/S0193-2306(04)10001-X

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis ? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cronin, M., Erkens, D. H., Schloetzer, J. D., & Tinsley, C. H. (2021). How controlling failure perceptions affects performance: Evidence from a field experiment. *Accounting Review*, *96*(2), 205–230. https://doi.org/10.2308/TAR-2018-0146

Das, R., Jain, K. K., & Mishra, S. K. (2016). Archival Research: A Neglected Method in Organization Studies. *Benchmarking: An International Journal*. https://doi.org/https://doi.org/10.1108/BIJ-08-2016-0123

Dyckman, T. R., & Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons*, 28(3), 695-712. https://doi.org/10.2308/acch-50818

Flammer, C., & Kacperczyk, A. (2016). The impact of stakeholder orientation on innovation: Evidence from a natural experiment. *Management Science*, *62*(7), 1982–2001. https://doi.org/10.1287/mnsc.2015.2229

Floyd, E., & List, J. A. (2016). Using Field Experiments in Accounting and Finance. *Journal of Accounting Research*, *54*(2), 437–475. https://doi.org/10.1111/1475-679X.12113

Forker, E., Grabner, I., & Sedatole, K. (2020). *Does learning by disaggregating accelerate learning by doing? The effect of forecast disaggregation on the rate of improvement in demand forecast accuracy*.

Frezatti, F., Aguiar, A. B., Wanderley, C. A., & Malagueño, R. (2015). A pesquisa em contabilidade gerencial no Brasil: desenvolvimento, dificuldades e oportunidades. *Revista Universo Contábil*, 11(1), 47-68. http://dx.doi.org/10.4270/ruc.2015147-68

Gassen, J. (2014). Causal inference in empirical archival financial accounting research. *Accounting, Organizations and Society*, *39*(7), 535–544. https://doi.org/10.1016/j.aos.2013.10.004

Haesebrouck, K. (2021). The Effects of Information Acquisition Effort, Psychological Ownership, and Reporting Context on Opportunistic Managerial Reporting*. *Contemporary Accounting Research*, *38*(4), 3085–3112. https://doi.org/10.1111/1911-3846.12712

Harrison, G. W. (2005), Field experiments and control. In: Harrison, G. W., Carpenter, J., & List, J. A. (Eds.) *Field Experiments in Economics* (*Research in Experimental Economics, Vol. 10*), Emerald Group Publishing Limited, Bingley, pp. 17-50. https://doi.org/10.1016/S0193-2306(04)10002-1

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Hesford, J. W., Lee, S. H., Van der Stede, W. A., & Young, S. M. (2006). Management Accounting: A Bibliographic Study. *Handbooks of Management Accounting Research*, *1*, 3–26. https://doi.org/10.1016/S1751-3243(06)01001-7

Hiebl, M. R. W., & Richter, J. F. (2018). Response Rates in Management Accounting Survey Research. *Journal of Management Accounting Research*, *30*(2), 59–79. https://doi.org/10.2308/jmar-52073

Ikäheimo, S., Kallunki, J.-P., University, S. M., & Schiehll, E. (2018). Do White-Collar Employee Incentives Improve Firm. *Journal of Management Accounting Research*, *30*(3), 95–115. https://doi.org/10.2308/jmar-51902

Kinney, W. R. (2019). The Kinney Three Paragraphs (and More) for Accounting Ph.D. Students. *Accounting Horizons*, *33*(4), 1–14. https://doi.org/10.2308/acch-52451

Laviers, L., Sandvik, J., & Xu, D. (2021). *CEO Pay Ratio Voluntary Disclosures and Investor Reactions*.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*(April), 19–40. https://doi.org/10.1016/j.jom.2018.10.003

Lourenço, S. M. (2019). Field Experiments in Managerial Accounting Research. *Foundations and Trends® in Accounting*, *14*(1), 1–72. https://doi.org/10.1561/1400000059

Luft, J., & Shields, M. (2002). Zimmerman's Contentious Conjectures: Describing the Present and Prescribing the Future of Empirical Management Accounting Research. *European Accounting Review*, *11*(4), 795–803. https://doi.org/10.1080/0963818022000047091

Luft, J., & Shields, M. D. (2014). Subjectivity in developing and validating causal explanations in positivist accounting research. *Accounting, Organizations and Society*, 39(7), 550-558. https://doi.org/10.1016/j.aos.2013.09.001

Mcvay, S. E. (2011). Discussion of Do Control Effectiveness Disclosures Require SOX 404(b) Internal Control Audits? A Natural Experiment with Small U.S. Public Companies. *Journal of Accounting Research*, *49*(2), 449–456. https://doi.org/10.1111/j.1475-679X.2011.00403.x

Moers, F. (2006). Doing Archival Research in Management Accounting. *Handbooks of Management Accounting Research*, *1*, 399–413. https://doi.org/10.1016/S1751-3243(06)01016-9

Mortensen, T., Fisher, R., & Wines, G. (2012). Students as surrogates for practicing accountants: Further evidence. *Accounting Forum*, *36*(4), 251–265. https://doi.org/10.1016/j.accfor.2012.06.003

Mucci, D. M., Frezatti, F., & Bido, D. de S. (2021). Enabling design characteristics and budget usefulness. *RAUSP Management Journal*, *56*, 38–54. https://doi.org/10.1108/RAUSP-04-2019-0058

Murphy, P. R., Wynes, M., Hahn, T.-A., & Devine, P. G. (2019). Why are People Honest? Internal and External Motivations to Report Honestly. *Contemporary Accounting Research*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Nascimento, A. R., Junqueira, E., & Martins, G. A. (2010). Pesquisa Acadêmica em Contabilidade Gerencial no Brasil: Análise e Reflexões sobre Teorias, Metododologias e Paradigmas. *Revista de Administração Contemporânea*, *14*(6), 1113–1133.

Oehlert, G. W. (2003). A First Course in Design and Analysis of Experiments. In *The American Statistician* (Vol. 57). https://doi.org/10.1198/tas.2003.s210

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Roe, B. E., & Just, D. R. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics*, 91(5), 1266-1271. https://doi.org/10.1111/j.l467-8276.2009.01295.

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). Research Methods for Business Students. In *Pearson* (8th ed., Vol. 3). https://doi.org/10.1108/qmr.2000.3.4.215.2

Shadish, W. R. ., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. In *Houghton Mifflin Company*. https://doi.org/10.1016/j.evalprogplan.2004.01.006

Smith, M. (2022). *Research Methods in Accounting* (6th ed.). SAGE.

Spekle, R. F., & Widener, S. K. (2018). Challenging Issues in Survey Research: Discussion and Suggestions. *Journal of Management Accounting Research*, *30*(2), 3–21. https://doi.org/10.2308/jmar-51860

Sprinkle, G. B., & Williamson, M. G. (2006). Experimental research in managerial accounting. *Handbooks of Management Accounting Research*, *1*, 415-444. https://doi.org/10.1016/S1751-3243(06)01017-0

Swieringa, R. J., & Weick, K. E. (1982). An Assessment of Laboratory Experiments in Accounting. *Journal of Accounting Research*, *20*, 56–101.

Trochim, W. M., Donnelly, J. P., & Arora, K. (2016). *Research Methods - The essential knowledge base* (2nd ed.). Cengage Learning.

Trottier, K., & Gordon, I. M. (2018). Students as surrogates for managers: Evidence from a replicated experiment. *Canadian Journal of Administrative Sciences*, *35*(1), 146–161. https://doi.org/10.1002/cjas.1377

Van der Stede, W. A., Young, S. M., & Chen, C. X. (2006). Doing Management Accounting Survey Research. *Handbooks of Management Accounting Research*, *1*, 445–478. https://doi.org/10.1016/S1751-3243(06)01018-2

Van der Stede, W. A. (2014). A manipulationist view of causality in cross-sectional survey research. *Accounting, Organizations and Society*, 39(7), 567-574. https://doi.org/10.1016/j.aos.2013.12.001

Wouters, M., & Wilderom, C. (2008). Developing performance-measurement systems as enabling formalization: A longitudinal field study of a logistics department. *Accounting, Organizations and Society*, *33*(4–5), 488–516. https://doi.org/10.1016/j.aos.2007.05.002