

Conducting Research with Quasi-Experiments: A Guide for Marketers

Avi Goldfarb and Catherine Tucker*

March 28, 2014

Abstract

This paper provides a procedural guide for marketing academics interested in applying quasi-experimental methods. We outline the various types of econometric methods, lay out an etiquette, and describe a number of examples of related research by marketing scholars. We discuss three steps: identifying the research question, outlining the identification strategy, and understanding the mechanism. We detail a variety of possible identification strategies, including difference-in-differences, regression discontinuity, and instrumental variables. For each, we emphasize the importance of clearly communicating the identifying assumptions underlying the assertion of causality.

Keywords: quasi-experiments, econometrics, practitioner's guide

*This paper builds on presentations we have given at the 2013 Workshop on Quantitative Marketing and Structural Econometrics and at the 2012 and 2013 ISMS Doctoral Consortia. We thank David Godes, Brett Gordon, Avery Haviv, Joon Ho Lim, Cristina Nistor, and Nathan Yang for comments.

1 Introduction

At the heart of much of applied statistics and econometrics is the equation:

$$y = f(\beta X, \epsilon) \tag{1}$$

Work with quasi-experiments uses this equation with a specific focus: Can a *single* covariate x in the vector of X be demonstrated to cause y ? This is the dominant empirical framework for published economics articles in labor, health, public finance, and innovation. Perhaps because of the historical focus of marketing science on building predictive models that are helpful for managers, it has been less the focus of academic marketing. However, there are many marketing research questions that ask whether a particular x causes a shift in y . Rather than provide the technical details that most marketing scholars received during their doctoral studies, the purpose of this paper is to guide marketing scholars on how to undertake and communicate quasi-experimental research in a way that is credible and easily understood.

Quasi-experimental tools mimic the random assignment that is inherent in lab experiments and that is often referred to as the ‘gold standard’ for identifying causal relationships. By using real world data, quasi-experiments often suffer from fewer external validity concerns compared to lab experiments. Manski (2007) (p. 137) notes that, “In principle, the argument [in favor of experiments] applies both to controlled experiments, in which a researcher purposefully randomizes treatment assignments, and to so-called natural experiments, in which randomization is a consequence of some process external to the research project.” The challenge in the quasi-experimental setting is to assess whether the quasi-experiment credibly proxies for random assignment.¹ Economists and statisticians have developed a

¹As in Meyer (1995) (p. 152), we use the term ‘quasi-experiment’ instead of ‘natural experiment’ to emphasize that “such studies are not quite experiments.”

well-established set of tools that enable researchers to identify potential quasi-experiments and then to assess the credibility of those experiments.

In this article, we provide a description of these tools, with an emphasis on the application of quasi-experimental tools to marketing problems. In doing so, we build on a large number of books and articles that have covered similar material for economics, policy, and sociology audiences. Much of the material overlaps with Angrist and Pischke (2009), Manski (2007), Meyer (1995), Imbens and Wooldridge (2009), and others. As such, we aim first to synthesize the ideas in this literature for a marketing science audience, in much the same way as the methods papers from the 2010 Structural Workshop described structural methods to the same audience (Gordon et al. (2011), Reiss (2011), Mela (2011), Chintagunta and Nair (2011), and Ellickson and Misra (2011)). Second, we aim to provide a guide to using the key techniques in the context of marketing problems to highlight their usefulness to marketing scientists who had not been previously exposed to such techniques. As a result, we hope to increase the understanding of these techniques for marketing scientists who do not plan to use these techniques, so that they have a sense of how to evaluate them.

A paper that successfully uses the quasi-experimental econometric approach answers the following three questions:

1. Research question: Do we care whether x causes y ?
2. Identification strategy: Does x really cause y to shift?
3. Mechanism: Why does x cause y to shift?

The first (and hardest) stage is identifying a question where marketing scholars actually care whether x causes y . It is hard because many of the y 's and x 's we can measure well are (unfortunately) uninteresting. Therefore, researchers who do quasi-experimental research do best if they start not with the data but instead start by asking themselves rhetorically - suppose I convincingly showed that a increase in x increases y - who would care?

The second stage is composed of two parts. The first part is working out a setting and dataset which makes it feasible to identify whether x causes y . This is challenging and requires both luck and perseverance. The second part is more straightforward and involves constructing tables and figures that convince a reader. We say that this is straightforward simply because there is a well-honed toolkit. The bulk of this article is dedicated to guiding readers through this toolkit. Underlying this is the recognition, description, and presentation of the identification strategy, or “the manner in which a researcher uses observational data (i.e. data not generated by a randomized trial) to approximate a real experiment” (Angrist and Pischke (2009), p. 7).

The third stage requires thoughtfulness about the underlying economic or behavioral mechanism. This stage uses elements of the toolkit from the second stage and combines them with theory to help understand the identified relationships. This is typically done by demonstrating that the effect identified in the second stage is larger in situations where theory suggests it will be large, and/or by demonstrating the effect identified in the second stage goes away in situations where theory suggests it should not apply. This serves two purposes. First, it provides further evidence that the effect identified in the second stage is likely to be causal. Second, it provides an understanding of the drivers of the effect.

It is important to note that, for many applications in marketing, causal identification of specific relationships does not matter. Prediction modeling has different objectives, and often does not require the researchers to tackle the identification problem directly (for examples, see Shmueli (2010)). The objective in prediction modeling is to assess which x in X , which structure of ϵ , and/or which functional form f , can be used to best predict y . It is often irrelevant to the research objective whether x causes y or y causes x or something else causes both, so long as a precise estimate of y can be gained out of sample. In contrast, identification should be a primary focus for researchers asking whether x causes y , where the objective is to get β . Quasi-experimental work aims to determine causal relationships

by identifying situations in which (under a clear set of assumptions) the variation in x can be seen as random for the purposes of assessing its impact on y .

To summarize, the best quasi-experimental papers identify an interesting research question. They then spend the bulk of the paper assessing the degree to which the quasi-experimental variation can be treated as true random assignment, clarifying the identifying assumptions behind the analysis. Clarity in the assumptions is important because quasi-experimental variation will necessarily lead to an imperfect assessment of the research question. The purpose is to get to a point where the reader is convinced that the results move our understanding of the research phenomenon forward. The remainder of this article is devoted to discussing this process in more detail.

2 Identification

2.1 Why the obsession with identification?

The problem of identification is that “many different theoretical models and hence many different causal interpretations may be consistent with the same data.” Therefore, “The justification for interpreting an empirical association causally hinges on the assumptions required to identify the causal parameters from the data” (Heckman (2000), p. 47).

One way to describe this issue is through the ‘potential outcomes approach,’ developed by Jerzy Neyman, Donald Rubin, and others (Rubin (2005)) provides a history of the ideas). For any discrete event/policy (x), each i has two possible outcomes:

- y_{i1} if the individual i experiences x
- y_{i0} if the individual i does not experience x

The difference between the two is the causal effect. The identification problem occurs because a single individual i cannot both receive the treatment and not receive the treatment. Therefore, only one outcome is observed for each individual. The unobserved outcome is

called the “counterfactual.” The unobservability of the counterfactual means that assumptions are required. The identification problem means that those who experience x , and those who don’t, are different in unobserved ways. Even with random assignment, the fundamental issue still holds, in the sense that the same individual does not experience the treatment and control conditions (at the same time).

Still, random assignment solves the inference problem as, *ex ante*, the “unobserved ways” should not matter. Therefore, random assignment is often called the gold standard of identification. List (2011) (p. 8), in his justification of the use of field experiments, argues that “The empirical gold standard in the social sciences is to estimate a causal effect of some action.” Angrist and Pischke (2009) (p. 11) emphasize that “The most credible and influential research designs use random assignment.” The reason for this is that, under random assignment, “the difference in outcomes across treatment groups captures the average causal effect” (p. 15). Paraphrasing Ronald Fisher, Manski (2007) (p. 136) notes that under random assignment, “the distribution of outcomes by members of a treatment group will be the same (up to random sampling error) as would be observed if the treatment in question were received by all members of the population.”

However, in many situations, experiments are not feasible, not appropriate, or too costly. In general, if we think of the ‘four Ps’ taxonomy, while experimentation is increasingly used to inform advertising decisions, practitioners and researchers rarely run field experiments to inform channels and product development because they are too time-consuming or often require a level of measurement of long-term implications that precludes feasible experimentation.² In addition, field experiments with pricing are often problematic in that customers can find them unfair. More generally, as Andrew Gelman writes, “Given the manifest virtues of experiments, why do I almost always analyze observational data? The short answer is

²Practitioners have gotten around these challenges in several ways. For example, conjoint analysis gives experimental variation, albeit in an artificial setting. Test markets enable some experimentation for new products, though often without a clear treatment/control structure.

almost all data out there are observational” (Gelman, 2010).

In quasi-experimental work, the researcher’s goal is to make the unobserved ways in which the treatment and control groups differ as untroubling as possible to the researcher and the reader. This goal is achieved through a clear and well-considered identification strategy.

2.2 Identifying a Quasi-Experiment

Angrist and Pischke (2009) (p. 7) define an identification strategy “to describe the manner in which a researcher uses observational data (i.e. data not generated by a randomized trial) to approximate a real experiment.” Therefore, the objective in such research of choosing an identification strategy is to find something that approximates random assignment to compensate for the non-viability of field experiments for many questions.

To do this, researchers look for sources of exogenous variation: A shock to the system that means that some individuals (the treated) are exposed to x , but not others. The variation could occur because of some random shock, or because the entity that decided whether people were exposed to x or not simply did not care about the outcome. Sometimes it is helpful to realize that the variation could happen at many levels: country, state, city, firm, establishment, street corner, individual, publication, website visit, invention, etc. The key considerations are that the reason for the variation needs to be understood, and that the relationship between the variation and the outcome of interest is driven only by the relationship between x and y and not by some other factor. Under the quasi-experimental framework, identification of parameters should not be driven by the functional form assumptions of the model. It is variation in the data that should drive the identification of the causal relationship.

We want to emphasize that rather than large and very visible shocks, it is often best to look for quasi-experimental variation in mundane events such as contract changes, state regulation, individual-level life changes, shifts in firm policy that did not occur because of

an anticipated effect on the outcome of interest, etc. We describe several examples below.

3 Quasi-Experimental Data Analysis

Once the researcher has found a setting that may help identify the causal effect of interest, the next steps involve exploring the raw data to see the degree to which the quasi-experiment is credible. In particular, the researcher should see that the treatment and control groups are similar in dimensions other than whether they received the treatment. Perhaps the ideal thought experiment here is Zhang (2010), whose treatment and control were a pair of kidneys from the same person. Most research settings are less favorable. Still, researchers should show a comparison of mean values of demographic characteristics and behaviors for the two groups. This comparison should support the argument that these groups are similar.

In cases where the treatment occurs in the middle of a time series, demonstrating that the treatment and control groups were similar prior to the arrival of the treatment can be a particularly powerful argument for exogeneity of the treatment. Many papers use a graph that shows that before the treatment occurred, the treatment and control groups were on a similar trend and had similar values; then, after the treatment occurred, the trajectory of the treatment group changed but not the control group.

After establishing similarity between the treatment and control groups in the raw data, the next step is typically to conduct regression analysis that demonstrates the effect of interest. Next, we discuss three different broad identification strategies using quasi-experiments and the process that authors should undertake to convince themselves and their readers that they have identified a causal effect.

3.1 Difference-In-Differences

A standard difference-in-differences (or ‘diff-in-diff’) analysis compares a treatment group and a (quasi-)control group before and after the time of the treatment. The ‘treatment’ is not true random experiments, but rather some ‘shock.’ Examples of such shocks include

Table 1: The Diff-in-Diff

	Treatment	Control
Before	A	B
After	C	D

government regulation (e.g. some states change their policy while others do not), and platform actions (e.g. one platform changes the information provided to users and another does not). In each of these cases the researcher needs to provide evidence that the shock can be seen as quasi-experimental for the purposes of answering the research question, and the identifying assumptions need to be clearly stated. Compared to a simple comparison (or single difference) analysis, difference-in-differences methods generate a baseline for comparison between the treatment and the control group. By looking at the change in the treatment group relative to the control group, difference-in-differences enables the researcher to control for many of the most obvious sources of heterogeneity across groups.

Table 1 shows the basic structure of a difference-in-differences data set. Under the assumption that the time trends for the treatment and control groups are the same, except for the treatment itself, the causal effect of the treatment on those in the treatment group is $(C-A)-(D-B)$.

For example, in Goldfarb and Tucker (2011), we examine the impact of privacy regulation on the effectiveness of online advertising. In late 2003 and early 2004, many European countries implemented new restrictions on how firms could collect and use online data. We used data on the success of nearly 10,000 online display advertising campaigns in Europe, the United States, and elsewhere between 2001 and 2008. We compared the change in effectiveness of the ad campaigns in Europe to the change in effectiveness of the ad campaigns outside of Europe. Thus, the first difference is the change in the campaign effectiveness $((C-A)$ and $(D-B))$, and the second difference is the change in Europe relative to elsewhere $((C-A)-(D-B))$. Compared to before the regulation, ad campaigns became 2.8 percentage

points less effective in Europe after the regulation. In contrast, compared to before the European regulation, ad campaigns became 0.1 percentage points more effective outside of Europe after the European regulation was implemented.³

While a difference-in-differences regression can be represented in a 2×2 table as in Table 1, it is typically analyzed with regression analysis in order to allow researchers to control for factors that may change over time and across individuals. The simplest version of this regression is:

$$y_{it} = \alpha_1 TreatmentGroup_i + \alpha_2 AfterTreatment_t + \beta TreatmentGroup_i \times AfterTreatment_t + \epsilon_{it} \quad (2)$$

where y is the outcome of interest, i represents the individual, t represents the time, and ϵ_{it} represents the error. The key focus of the difference-in-differences specification is on β which captures the effect of the crucial interaction term. Usually, researchers add controls X_{it} to address additional omitted variables concerns, such as an observed covariate that may not affect that treatment and control groups in the same way:

$$y_{it} = \alpha_1 TreatmentGroup_i + \alpha_2 AfterTreatment_t + \beta TreatmentGroup_i \times AfterTreatment_t + \gamma X_{it} + \epsilon_{it} \quad (3)$$

The above specification implies a repeated cross section. In Goldfarb and Tucker (2011),

³Difference-in-differences methods do not require a time series component. For example, in their study of the impact of Quebec's ban on advertising to children, Dhar and Baylis (2011) compare the difference between francophone and anglophone households inside Quebec (where francophone households were affected by the ban) to the difference between Canadian francophone and anglophone households outside of Quebec (where the ban did not apply).

one potential concern with such a specification is that the campaigns might have changed differently in Europe relative to elsewhere for reasons other than the regulation, something we ruled out with mechanism checks that we discuss in section 4 of that paper.

When researchers have access to a panel, it is possible to address this concern directly by observing the same individuals, or the same campaigns, both before and after the timing of the treatment. It is then possible to add fixed effects to control for all individual-level (time-invariant) heterogeneity. Furthermore, if the data includes more than two time periods, then adding time-specific fixed effects controls for all time-period specific heterogeneity (across all individuals). With individual and time fixed effects, the difference-in-differences regression is:

$$y = \beta \textit{TreatmentGroup}_i \times \textit{AfterTreatment}_t + \gamma X_{it} + \mu_i + \tau_t + \epsilon_{it} \quad (4)$$

where μ_i is the individual-level fixed effect and τ_t is the time-period fixed effect. The fixed effects mean that the main effect of $\textit{TreatmentGroup}_i$ and $\textit{AfterTreatment}_t$ drop out because they are collinear with the fixed effects. If possible, it is often desirable to difference out, rather than estimate, the fixed effects to avoid bias due to the incidental parameters problem (e.g. Lancaster (2000)). Most standard statistical packages automatically condition out the individual fixed effects from fixed effects panel data models where possible.⁴

A final tweak on this model is that if the treatment occurs at different times, meaning that individuals are treated at different times, then the $\textit{AfterTreatment}$ variable can change with i and t . For example, Chevalier and Mayzlin (2006) study how a book review posted

⁴For example, the fixed effects specification of Stata's `xtreg` function uses differences from average values. The fixed effects specifications of Stata's `xtlogit` and `xtpoisson` also condition out the individual-level fixed effects.

on Amazon affects sales of that book on Amazon, compared to sales of that book at barnesandnoble.com. Different books are reviewed at different times. Therefore, the ‘treatment’ here is the review a book receives and the *AfterTreatment* period occurs at different times for different books.

A key issue in difference-in-differences analysis is correlated errors in observations because the outcome is often observed at a finer level than the treatment. For example, the researcher might observe treatment and control groups for a number of advertising campaigns over a long time period. For each campaign, the researcher might have data on many individuals per campaign and many time periods per individual. It is important to recognize that the choices of the same individual in many time periods are likely to be correlated. Bertrand et al. (2004) emphasized that failure to control for the correlation between these choices will lead to an overstatement of the effective degrees of freedom in the data and therefore standard errors will be biased downwards. They suggest clustering standard errors by individual over time to address this issue and provide Monte Carlo evidence that clustering is likely to lead to robust inference.

Similarly, Donald and Lang (2007) emphasize that if individual responses to the same treatment are likely to be correlated (for example, because of close physical or social proximity), clustering standard errors by groups of individuals is a conservative way to estimate standard errors. Researchers often need to decide on the size of the clusters. For example, in studying ready-to-eat breakfast cereals, is the correct unit the company (e.g. General Mills), the brand (e.g. Cheerios), or the subbrand (e.g. Honey Nut Cheerios)? The answer depends on the data and research question. Each cluster should contain those observations most likely to be correlated with each other.⁵

⁵A handful of recent papers have recognized the challenges in applying clustered standard errors (which rely on consistency arguments and large samples) to a small number of clusters. With a small number of clusters, a correction is needed such as the ‘wild bootstrap’ developed by Cameron et al. (2008). A key choice is then to design the quasi-experimental setting to ensure that there are a sufficient number of clusters for inference, but that each cluster captures the most closely related observations.

To summarize and expand on some of the issues above, we think it is useful to list a ‘diff-in-diff etiquette’:

Step 1: Explain and defend the “experiment.”

In the absence of random assignment, it is not possible to prove that the difference between the treatment and control group is exogenous. The purpose of this step is to clearly lay out the identification strategy, describe the assumptions behind that strategy, and explain why those assumptions are reasonable in this situation. Examples of recent difference-in-differences approaches by marketing researchers include:

1. Contract disagreements or changes: Chiou and Tucker (2012) use a removal of news content that resulted from a breakdown in contract negotiations between the Associated Press and Google to study the effect of aggregators on downstream news websites.
2. Shifts in firm policy: Zentner et al. (2013) use brick-and-mortar video store closings to identify differences between online and offline purchasing patterns.
3. Individual-level life changes: Bronnenberg et al. (2012) use consumer migration to new locations as a quasi-experiment to study the causal impact of past experiences on current purchases.
4. Regulatory Changes: Tucker et al. (2013) use a change in Massachusetts regulation of home sale listings to identify the effect of information about time on the market on house prices. Qian (2014) uses a change in counterfeit enforcement in China that affected some locations more than others to identify the impact of counterfeits on sales.

It is also important to consider whether the quasi-experiment affects x in an interesting way. For example, firms might be interested in the marginal effect of advertising on sales, while a quasi-experiment using difference-in-differences might involve completely shutting down all advertising. More broadly, it is important to remember that no quasi-experiment

used for a difference-in-differences will be perfect. Even experiments that rely on random assignment by researchers have their flaws. One of the most important aspects of the identification strategy is to clearly lay out the identifying assumptions.

For example, Busse et al. (2006) (p. 1261) use week-to-week variation in promotions offered to retailers by car manufacturers to assess the degree to which retailers pass such promotions through to manufacturers. They write that “The identifying assumption...is that prices of cars in the same segment that are not on promotion in a given week are a valid counterfactual for the prices that would have been obtained on the promoted car in the absence of a promotion.” In our study of the impact of European privacy regulation on the effect of online advertising (Goldfarb and Tucker, 2011) (p. 65), we describe “[...]ur fundamental identifying assumption that the European campaigns and the European respondents do not systematically change over time for reasons other than the regulations.” In both papers, a substantial part of the empirical effort is dedicated to exploring the validity of these assumptions.

In addition to a written explanation and defence of the identifying assumptions, researchers can do further analysis to support their argument. Doing so involves completing steps two through eight of the difference-in-differences etiquette, and paying careful attention to mechanism checks (discussed in Section 3 below). Unfortunately, however, there is no formula for a convincing explanation and defense of the experiment. Except in cases of random assignment, it is not possible to prove that the identifying assumption is right. Instead, the objective for the authors is to pursue projects only when they can convince themselves (and their readers) that the causal interpretation is more plausible than other possible explanations.

Step 2: Present the raw data in terms of a graph.

Show the outcomes of interest for the treatment and control group before and after the treatment. This communicates the basic variation in the data that will be used to identify the effect in the regression analysis.

Step 3: Show treatment and control groups are similar pre-treatment in terms of their observable characteristics.

This helps defend the identification strategy. If the treatment and control groups are substantially different in the pre-treatment, the control group is unlikely to be a good proxy for the counterfactual and the quasi-experiment is unlikely to be valid.

Step 4: Present baseline estimates. Cluster standard errors as appropriate.

This typically appears in the form of a regression table with several different specifications. For example, the first column might not include any controls beyond the fixed effects, and the next set of columns might add controls. Comparing the coefficient of interest in the models with and without controls is informative about how big the impact of the omitted variables has to be relative to the observed controls in order for an omitted change over time in the treatment group to drive the result (Altonji et al., 2005).

Specifically, Altonji et al. (2005) provide a method to compare the role of the included and omitted variables. The method allows the researcher to examine how much the effect of interest changes as controls are added and then to assess how important the omitted variables would have to be for the treatment effect to go away. They then provide details on how to more formally assess how big and correlated the effect of the omitted variables needs to be relative to the controls.⁶ While the formal method is useful, many researchers (e.g. Mayzlin et al. (2012), Anderson et al. (2013)) leverage the more basic insight that there is information in the impact of the controls on the measured effect of interest. This does

⁶The method has been applied and extended in the marketing literature by Shin et al. (2012).

not mean that results are invalid if the controls do change the estimated effect substantially, however, indeed identifying relevant omitted variables can help provide further support for the causal interpretation.

As discussed above, in presenting the estimates, standard errors should be clustered at the appropriate level in order to ensure the results do not exaggerate the statistical power of the coefficients (Bertrand et al. (2004), Donald and Lang (2007)).

Step 5: Investigate pre-treatment patterns.

It is useful to show that behaviors were similar in the period prior to the policy change across the treatment and control groups. This reinforces the results from Steps 2 and 3. This is often done by estimating β for each period in the data and plotting out the estimates. In particular, replace $\beta \text{ TreatmentGroup}_i \times \text{AfterTreatment}_t$ with a series of covariates for each time period before and after the treatment.⁷

For example, Qian (2008) studies the effect of counterfeiting on prices of authentic footwear in China. Figure 1 reproduces figure V from her paper. It shows the year-by-year coefficients of a regression of prices on counterfeiting entry. The figure demonstrates clearly and visually that there is a sudden change at the time of entry, that the before and after periods are different, and that the values of the treatment and control group are similar prior to the time of entry (the coefficients are near zero before entry). Even though entry occurred in different years for different brands, by defining the horizontal axis relative to the year of entry, many different treatments can be viewed using the same scale.

Step 6: Conduct multiple robustness checks.

The specific robustness checks chosen will depend on the exact context. With electronic appendices and generally cheap computation, it is possible to show robustness to a large number of alternative specifications. Here empirical work with quasi-experimental methods

⁷Sometimes, researchers present just the raw data (Step 2) or just the regression-based graph (Step 5) if the graphs are sufficiently similar.

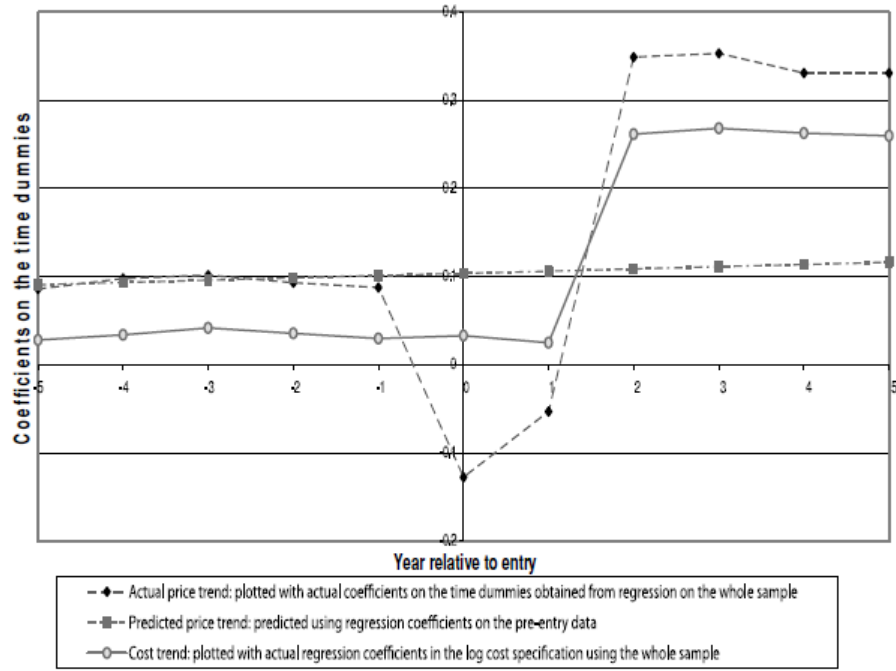


FIGURE V
Time Plot of Average Log Deflated Authentic Prices and Costs
Note. This figure plots the regression coefficients on the dummies (indicating the year relative to counterfeit entry), with log deflated authentic price or cost as the response variable.

Figure 1: Figure V from Qian (2008)

differs substantially from research using forecasting models. The aim is not to show one specification (or model) and defend it. The idea is to try to show that the estimate of β remains broadly consistent across a vast range of possible models.

Here are some examples of useful robustness checks:

1. Different choices of controls, including no controls at all.
2. Different functional forms. For example, if using a linear probability model show robustness to logit and probit.⁸
3. Different choices of the time period under study. Researchers often can choose when to start and end the sample. For example, for a treatment that occurs in 2004, it is partly the researcher's choice whether the period under study should be 2002 to 2006, 2000 to 2008, 1995 to 2013, 1900 to 2013, etc.
4. Different dependent variables. There might be several different but related dependent variables that relate to the outcome of interest.
5. Different choices of the size of the control group. Researchers choose whether all the data should be used in the control group, or only a subset of the data that is 'close' to the treatment group (for example as measured by a propensity score). Researchers can also choose how to define the treatment group.

It is unlikely that every robustness check will yield the same level of significance as the initial specification. Researchers (and reviewers) should not expect every specification to yield the exact same results. The key is to communicate when the results hold up and when

⁸The choice between linear probability models and non-linear models such as logit is widely debated. Angrist and Pischke (2009) argue for linear probability models because they are simple to interpret and consistent under a basic set of assumptions. Others argue against them because they are inefficient (and inconsistent if the assumptions are violated). In cases like this, where the literature does not give clear guidance on the choice of model, showing robustness to different choices is optimal.

significance fails. This will consequently help inform the reader what drives the statistical power behind the results.

Step 7: Discuss the assumptions behind homogeneous treatment effects and/or explain why the treated population is inherently interesting.

This step bounds the external validity of the analysis. It discusses the assumptions required for the analysis to capture the average treatment effect, rather than a more local effect which is an artifact of the data sample or the source of quasi-experimental variation.

One reason a research setting may fail to be externally valid is if the treated population is unrepresentative. It is therefore important to consider whether the measured effect is specific to the sample affected by the shock or whether it will apply more generally. As with real experiments, quasi-experiments are best when the variation in x is closely related to the general question of interest. If the available data are too narrow to inform the general research question, then the causal effect identified in the quasi-experiment may not be useful. There is a lengthy literature that emphasizes which types of estimates are relevant to which research questions, demonstrating that estimates of the average treatment effect, the average treatment effect on the treated, and the local average treatment effect can differ.⁹

More generally, it is important to understand the underlying assumptions behind any broad interpretation of quasi-experimental results. The treatment effect is likely to be heterogeneous across places, across institutions, across time, and across demographics. While, in the presence of the right data, these heterogeneous treatment effects can provide an opportunity to identify the mechanism, heterogeneous treatment effects limit the recommendations and policy implications that arise from a difference-in-differences because data constraints usually push toward assuming more homogeneous treatment effects than is likely true.

⁹The treatment effect of most general interest is called the “average treatment effect” (ATE). This is the average effect of the treatment on all units. The “average treatment effect on the treated” (ATT) describes the effects on the treated units. The “local average treatment effect” (LATE) describes the effect of the treatment on the subpopulation that is induced by the treatment to change behavior.

To give a concrete example of these concerns embodied in the concept of a Local Average Treatment Effect (LATE), imagine a study that focused on the question of how mailing people multiple coupons affected purchases. The example ‘quasi-experiment’ is that sometimes the person responsible for data input misspelled names and consequently some consumers turned up twice in the database and received two coupons. One concern around this would be that if there were data inputting errors in the name, their address could be more likely to be misspelled. This would lead people who were supposed to be in the treatment group actually also to be systematically more likely to not be treated at all. Another concern would be what happens if the population whose names were misspelled were mostly of (say) Polish ancestry. Could that ancestry also lead to different and non-representative types of purchase behavior? Finally, it is important to question whether it is interesting to study an instance where the consumer received multiple coupons on the same day - surely most multiple coupon strategies would stagger their release?

An important aspect of the best practice regarding the external validity of results is to clearly lay out the assumptions and limitations. For example, Sun and Zhu (2013) use a quasi-experiment and difference-in-differences to examine the impact of advertising revenue on the type of content posted on Chinese blogs. While it might be tempting to interpret the results as suggestive of a broader impact of commercial interests on media, they are careful to emphasize the many differences between blogs and other media, between China and the rest of the world, and between the way the bloggers were compensated and other online advertising models. In this way, the paper explicitly limits the temptation of the reader to extrapolate too much.

An alternative solution is to use a difference-in-differences approach to identify relationships such as elasticities and then to use a structural model to identify the counterfactual of interest. In these cases, quasi-experimental methods serve as a complement for, rather than a substitute to, structure. For example, Anderson et al. (2013) use quasi-experimental

methods to identify the impact of the automotive brand preferences of parents on the brand preferences of their children. They then use structural methods to estimate the implications for firm strategy. Einav et al. (2010) use quasi-experimental variation in health insurance prices to identify price elasticity and then combine this measure with a structural model to estimate the welfare implications of adverse selection. Chung et al. (2014) use quasi-experimental variation around set quotas to identify the relationship between commissions and sales, and then use this variation in a structural model to determine optimal compensation schemes.

Step 8: Apologize for all that is still unproven and give caveats.

Any identification strategy relies on a set of assumptions. These assumptions need to be explicit throughout the paper. There are always some tests that cannot be run, for example due to lack of data. There are always some robustness checks that are weaker than others. There are always some steps from data to interpretation. While apologies do not mean all is forgiven, the objective should be to clarify the boundaries of the claims. Obfuscation is much worse than a clear summary of the identifying assumptions.

Busse et al. (2006) are particularly clear about the assumptions underlying their estimation. Their paper includes difference-in-differences and regression discontinuity results. For the difference-in-differences, as mentioned above, they state (p. 1261), “The identifying assumption of the difference-in-differences approach is that the prices of cars in the same segment that are not on promotion in a given week are a valid counterfactual for the prices that would have been obtained on the promoted car in the absence of a promotion.” This clearly lays out the key assumption. They then go on to explain, “Although we cannot observe this directly, we can examine the trends of promoted and nonpromoted cars in the period just prior to the promotion. If the trends are similar between cars that are soon to be promoted and cars that are not, that gives some assurance that the nonpromoted cars

are a valid counterfactual in the promotion period.” In our view, this is a particularly clear example of how researchers can explain the identification strategy, describe what they do to address it, and also detail what they cannot do.

Overall, difference-in-differences is a powerful tool for helping to identify causal effects. It can enable researchers to control for time-invariant individual-level heterogeneity, relying on the assumption that differences in the changes the treatment and control groups experience over time are driven by the impact of the treatment.

3.2 Regression Discontinuity

Regression discontinuity (RD) is another useful quasi-experimental technique in which the ‘experiment’ relies on an exogenous arbitrary threshold. As Imbens and Lemieux (2008) put it, “The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of the predictor being on either side of a fixed threshold.”

Hartmann et al. (2011) emphasize the promise of regression discontinuity as an identification strategy for marketing scholarship. They argue that many marketing interventions are based on thresholds of real or expected consumer behavior. For example, direct mail companies use the scoring policies for Recency, Frequency, Monetary (RFM) models. Consumers just above and just below the cutoff should be similar in many dimensions and their outcomes can be compared to assess the impact of the different mailings.

Another example is government policies based on firm size. For example, many government policies regarding requirements for firms to post calories, undertake layoffs, and provide benefits depend on the number of employees or other measures of firm size. By comparing firms just above and just below the threshold, it is possible to assess the impact of the policies on firm behavior.

It is also possible to use a regression-discontinuity argument to justify examining a short

time window around a policy change. In such cases, the effective source of the regression discontinuity is time. This is an alternative interpretation of the period-by-period regressions described above in section 3.1 on difference-in-differences. In general, in such cases it is still desirable to have some baseline control group. The other issue this kind of approach faces is that often the implementation and effective date of a policy change (especially if the policy is not digital) is staggered and subject to delays, making a clean discontinuity problematic.

As with difference-in-differences, there is also a sequence of regression discontinuity etiquette that the best papers using this technique tend to follow. Given that the methods have similar objectives, the etiquette overlaps.¹⁰

Step 1: Explain and defend the “experiment.”

Particular attention should be devoted to the source of the threshold and providing evidence that the threshold is essentially arbitrary and not likely to be linked to underlying discontinuities in behavior. Any discontinuity in the effect is assumed to be due to the treatment. The key assumption is that although the predictor may be correlated with outcomes, the association is assumed to be smooth.

This assumption is not always innocuous. Consider a \$50 cutoff for receiving a marketing incentive. If the firm promotes the threshold and consumers try to achieve it, then there might be a substantial difference between people who spend \$49 and people who spend \$51. Those who spend \$49 are likely to be unresponsive to the incentive because they did not try to cross the threshold in order to get the incentive. In contrast, those with exactly \$50 in spending might have selectively chosen to spend exactly enough in order to get an incentive that they planned to use. It is important to address the potential for such concerns directly.

¹⁰Imbens and Lemieux (2008) and Hartmann et al. (2011) provide further details on the identifying assumptions and methods.

Step 2: Present the raw data.

Provide graphical evidence for this discontinuity. For example, a recent debate in economics centered around the use of birth weight thresholds to determine the impact of intensive care units on medical outcomes for newborns. In the initial study, Almond et al. (2010) used the fact that birth weight threshold of 1.5kg is used to determine whether or not the newborn receives intensive medical treatment. In a critique of this work, Barreca et al. (2011) show however, that the children are placed just at the cut-off seem to have significantly worse outcomes than babies on either side of the cut-off. This is evidence against use of this discontinuity for identification. Barreca et al. (2011) state, “This may be a signal that poor-quality hospitals have relatively high propensities to round birth weights but is also consistent with manipulation of recorded birth weights by doctors, nurses, or parents to obtain favorable treatment for their children.”

Step 3: Show groups below and above the threshold are similar based on observables.

As in a difference-in-differences analysis, this helps defend the identification strategy. If the two groups are substantially different, the group below the threshold is unlikely to be a good proxy for the counterfactual of the group above the threshold and the quasi-experiment is unlikely to be valid.

Step 4: Present baseline estimates. Cluster standard errors as appropriate.

These will probably be very simple since there is typically just one discontinuity being estimated.

Step 5: Conduct multiple robustness checks.

As described above in the difference-in-differences analysis, there are several choices that go into the final regression specification that are not dictated by the identification strategy, by

theory, or by guidance from the literature. Showing robustness to functional form, covariates excluded, outcome used, and other choices will improve the credibility of the estimates.

Step 6: Discuss the assumptions behind homogeneous treatment effects and/or explain why the treated population is inherently interesting.

A limitation of regression discontinuity is that the results directly apply only to populations around the threshold. For example, comparing the \$49 spend with the \$51 spend may be informative about the impact of the marketing incentive on consumers who spend around \$50; however, consumers who typically spend a lot more or a lot less might be different.

Step 7: Apologize for all that is still unproven and give caveats.

Again, it is always important for researchers to lay out the identification strategy, assumptions, and limitations of the analysis.

Regression discontinuity can be a useful tool for helping to identify causal effects when there is a fixed threshold that determines whether a treatment occurs. While such situations are rare (besides time-based discontinuities), when these thresholds exist they can provide convincing evidence suggesting causal effects.

3.3 Instrumental variables

The basic idea behind using instrumental variables is that the covariate of interest contains both useful variation (to identify the causal effect of interest), and less useful variation (that confounds the effect). A good instrument is strongly correlated with the useful variation but uncorrelated with the confounding variation. In other words, one only uses the variation in x which can be explained by the plausibly exogenous shifter z . The standard two-stage model is estimated as:

$$x_1 = \gamma z + \theta X + \eta$$

$$y = \beta \hat{x}_1 + \phi X + \epsilon$$

The identification of the effect of x_1 on y relies on the ‘reduced form’:

$$y = \beta \gamma z + \hat{\phi} X + \epsilon$$

Therefore, from the quasi-experimental point of view, an instrumental variable can be seen as a treatment that affects the endogenous covariate directly.¹¹ This means that directly regressing the outcome of interest on the instrument (in one stage) will get the causal effect of interest, but it will not be properly scaled. The purpose of implementing two stages is to scale the treatment effect properly.¹² Using two stages allows the researcher to disentangle β and γ . In other words, two stages are needed to get the elasticity right, but the experiment happens at the level of the instrument and so, even though the focus is on the relationship between x and y , the intuition on causality happens at the level of the relationship between z and y .

Instrumental variables can be a less transparent solution to identifying causal effects than the techniques described above. Though the underlying mathematics of instrumental variables, regression discontinuity, and difference-in-differences are similar, it is sometimes harder to visualize how the quasi-experimental variation works in instrumental variables. There are three sources for the difficulty in implementing instrumental variables results in a transparent manner. First, in contrast to the binary nature of the exogenous variation in difference-in-differences and regression discontinuity, instruments are often continuous. This makes it more difficult to communicate the intuition for why the variation is exogenous to

¹¹While this perspective is somewhat different than the simultaneous equations perspective used in many introductory explanations, it is important to remember that the underlying mathematics is identical. The difference is in the intuition that helps identify useful instrumental variables and argue for their validity.

¹²There are many ways of operationalizing instrumental variables and this can be a place for highly technical tools. We emphasize the most simple two-stage least squares approach, but the intuition behind the role of instrumental variables as an identification strategy remains regardless of functional form assumptions.

the potential for omitted variables or simultaneity.¹³ Second, researchers have increasingly recognized the challenges of using ‘weak’ instruments. Essentially, instrumental variables techniques are consistent but biased and this bias can matter even in seemingly large samples (Stock et al., 2002). Weak instruments can lead to incorrect inference in which the bias of the weak instrument dominates the potential bias of the omitted variables. Of course, there can be weak quasi-experiments using difference-in-differences, but there the weakness is easier to spot when plotting out the raw data. Third, many researchers present instrumental variables results with different tests and with different norms. This makes it difficult to read and assess the validity of papers with instruments. Angrist and Pischke (2009) (p. 212-213) provide a sequence of steps to follow in an attempt to standardize practice. In light of the challenges to generating convincing instruments, the etiquette becomes particularly important. Given the similar objectives of difference-in-differences methods, regression discontinuity methods, and instrumental variables methods, there is substantial overlap in the etiquette. We summarize Angrist and Pischke’s etiquette as:

Step 1: Explain and defend the “experiment.”

The exclusion restriction is the cornerstone of any instrumental variables analysis. The exclusion restriction states that the instrument only affects y because it affects x . However, it is not possible to directly prove the validity of the exclusion restriction. Therefore, the validity of a given instrumental variables strategy necessarily depends on the rhetoric used to explain why the exclusion restriction might be expected to hold. The purpose of this step is to clearly lay out the identification strategy, describe the assumptions behind that strategy, and explain why those assumptions are reasonable in the situation at hand.

For example, in Shriver et al. (2013) they use the speed of wind as an instrument to

¹³The ability to use continuous instruments can also be seen a strength of instrumental variables techniques. It enables a more flexible set of counterfactuals because there are more treatments observed and used in the analysis. For example, while a discrete quasi-experiment on retailer discounts would allow the researcher to compare the impact of a small set of retailer discounts on sales, a continuous instrument for the discounts might allow the researcher to compare a variety of smaller and larger discounts.

provide an exogenous driver of posting to a user-generated content site about windsurfing. This allows them to understand the relationship between content creation and the creation of social ties. The argument for the exclusion restriction is that there is no other way plausibly that wind could affect the creation of social ties except through content creation. As they mention in the paper, plausible challenges to this exclusion restriction are that windy days could affect friendship formation directly because users meet future online friends at more windy surf locations. To deal with such challenges, the researchers present empirical data to suggest that the social ties that are being formed do not seem to reflect geography.

Another example is Lambrecht et al. (2011), who studied the effect of delays in the early part of a banking technology adoption process on ultimate usage. As an instrument that provides a source of exogenous variation in delays, they exploit the fact that Germany has a highly regulated system of public holidays and vacations, that vary at the state level to prevent freeways from becoming overly congested. This leads to delays in technology adoption in that particular period to customers in one state, and not in others. The exclusion restriction is that there is no other reason that vacations or public holidays in the few days surrounding adoption would affect ultimate usage except through delaying the ability to navigate the security protocols required to sign up for the online banking service. One challenge for the exclusion restriction could be that individuals who sign up for a banking service around public holidays are somehow systematically different in terms of their laziness or motivation from others. To counter this challenge, the researchers put forth evidence about the observable characteristics of users, showing that they are not different along any observable dimension.

Step 2: Test for power and overidentification.

Given that inference with weak instruments can be biased, there are a number of steps that researchers can take to assess whether the instruments have power:

1. Report the first stage. Assess whether the signs and magnitudes of the coefficients make sense.
2. Report the F-statistic on the excluded instruments. This helps determine whether the instruments are weak. Stock et al. (2002) advise that F-statistics below 10 suggest weak instruments, though, as Angrist and Pischke (2009) (p. 213) put it, “obviously this cannot be a theorem.”
3. If there are multiple instruments, report the first and second stage results for each instrument separately (at least in the appendix) because bias is less likely if there is only one instrument. Showing the results separately also helps the reader understand the intuition behind the quasi-experiment underlying each instrument. In addition, show the overidentification test to ensure that different instruments use different variation in increasing the apparent power of the analysis.
4. Conduct a Hausman test comparing OLS and instrumental variables. If the results change, reflect on whether they change in a direction that makes sense given the power of the instruments.
5. Compare the two-stage least squares results with the limited information maximum likelihood (LIML) results. If they are different, there may be a weak instruments problem.

Step 3: Do a reduced-form regression of the dependent variable (2nd stage) on the instruments.

Show the reduced form result of regressing the outcome directly on the instrument. Because this is an OLS regression, it is unbiased. This regression rarely appears in published papers (though perhaps it should). At the very least, the researcher should be confident that the instrument (z) has the expected direct effect on the outcome (y).

Step 4: Conduct multiple robustness checks.

Researchers have many choices, and it is important to show that the results are not driven by any particular choice.

Step 5: Discuss the assumptions behind homogeneous treatment effects and/or explain why the treated population is inherently interesting.

As above, this step bounds the external validity of the analysis. It discusses the assumptions required for the analysis to capture the average treatment effect, rather than a more local effect which reflects the data sample studied or the source of variation for the instrument.

Step 7: Apologize for all that is still unproven and give caveats.

The identification strategy relies on a variety of assumptions, some of which cannot be tested. It is important to make those assumptions and any limitations clear. As before, obfuscation is much worse than a clear summary of the identifying assumptions.

3.4 Summary of identification strategies

Each of the identification strategies summarized here has its strengths and weaknesses. Often, it is useful to combine different approaches in the same paper. For example, Busse et al. (2006) use both difference-in-differences and regression discontinuity (in time) to examine how much retailers pass manufacturer promotions through to end customers. And Qian (2008) combines a difference-in-differences strategy with counterfeit entry as the treatment with a convincing and high powered instrument on government regulation.

So far, we have focused on the identification strategy. Next, we turn to the other set of analyses necessary for a convincing quasi-experimental paper: The mechanism check.

4 Mechanism Checks

This discussion of the identification strategy has focused on the question “Does x cause y to shift?” Answering this question is key, but the most effective papers typically do not stop

there. After identifying a likely causal effect, it is important to assess why x causes y to shift. This is useful for two reasons. First, understanding underlying mechanisms is a key goal of social science. Second, mechanism checks often help make causal identification more convincing.

The first form of mechanism checks are falsification checks. These involve identifying what other groups would be affected by potential sources of bias that would not display the causal effect of interest. In this way, these can be seen as further robustness checks.

However, these falsification tests also provide an opportunity to help explore the behavioral mechanism. If the effect goes away when theory suggests it should, then this helps identify why it happens. For example, Lee and Bell (2014) showed that neighborhood social capital mediates the relationship between social influence and purchasing for highly visible goods such as fashionable apparel but that it does not mediate the relationship for diapers, which are hidden from view. Anderson et al. (2010) showed that sensitivity to sales tax was reduced when there were also sales signs - which is suggestive that sensitivity to sales taxes online is part of general patterns of price search online.

Having used this approach to help rule out any lingering concerns about identification, the researcher can then use a similar approach to provide suggestive positive evidence about the behavioral mechanism. If the effect is larger when theory suggests it should be, then this helps identify the mechanism. A simple approach is to estimate the effect separately by whether an individual is a member of a group which theory suggests should experience a bigger effect. Formal testing whether the difference is statistically significant, however, will require a three-way interaction between x , the source of variation, and group membership.

For example, after showing the European privacy regulation hurt online advertising, in Goldfarb and Tucker (2011) we first ran a falsification check, demonstrating that European consumers behaved like Americans when visiting American websites and that American consumers behaved like Europeans when visiting European websites. Then we explored

the mechanism and showed that the regulation especially hurt unobtrusive advertising and advertising on general interest websites, two situations where using data to target advertising is particularly valuable.

Overall, the mechanism check is important because it helps support claims of causal inference and because it enhances the likelihood that a paper is remembered. It is often not enough to measure something better. A paper is more likely to be remembered for the evidence that shown in support of a theory explaining why the result holds.

5 What happens if there is no quasi-experiment?

Sometimes, there is no quasi-experiment in the data. The question then is whether controls are sufficient to deal with the omitted variables problem. Adding controls through multiple regression addresses potential bias in the treatment effect by including covariates that are correlated with the treatment and the outcome in a linear way. As typified by Wooldridge (2000), a realistic perspective for such an approach is that “we can hope to infer causality.” Caution is warranted for two reasons. First, the linear multiple regression model assumes a particular functional form. Second, and more importantly, it is not possible to know whether the controls capture all of the relevant omitted variables.

Matching estimators help address the first concern. A matching estimator compares the outcomes of a treatment group and an artificial control group, where the treatment and controls groups are ‘matched’ based on similarity on observed characteristics (Todd, 2008). Thus, rather than assuming the linear structure, matching estimators allow for a non-parametric (i.e. flexible) relationship for controlling observables. If outcome measures are costly to obtain, matching saves time and effort in the data collection process.

Matching estimators are often mentioned as a solution to the potential outcomes problem. This is not quite accurate. It is true that matching estimators allow for flexible controls for observables; however, matching estimators do not address the second (and more substantive)

concern in using multiple regression to infer causality – that it is not possible to know whether the included covariates capture all of the relevant variables (Smith and Todd, 2008; Angrist and Pischke, 2009).

There are a variety of different matching estimators. What they have in common is the ability to identify similar individuals based on observables in a flexible way. Typically, this involves a ‘propensity score,’ which is a statistical prediction of how likely an individual is to be in the treatment group.

In the absence of the quasi-experimental variation which is the focus of this guide, researchers can at least derive a measure of how large the omitted variables bias has to be in order to change the conclusions. As mentioned above, Altonji et al. (2005) provide a method to assess how big the omitted variable bias has to be relative to the included controls in order for the documented results to go away. In this way, when the most obvious confounds have little impact, it provides a limited way to assess the plausibility of the causal interpretation even in the absence of a quasi-experiment. This strategy can be enhanced by showing multiple data sets with different potential biases yield similar results.

In some cases in the absence of a quasi-experiment, the use of such techniques are sufficient. In particular, if the researchers do not have reason to expect reverse causality, if the researchers have included a large number of controls (including the most obvious confounds), and if these controls do not change the estimated treatment effect, then it is reasonable for the researchers to clearly state the assumptions behind the interpretation and move to exploring the mechanism.

6 Conclusion

In this paper, we have presented the logic behind the quasi-experimental framework. In light of that logic, we have detailed several of the most commonly used methods for identifying causal effects. One article cannot comprehensively cover all relevant topics in detail, and

we have not attempted a comprehensive review of the marketing literature that uses these methods. Instead, we have used examples from the literature to shed light on the usefulness of these methods to marketing scholars and to provide a set of guidelines for their effective implementation. The objective of a quasi-experimental research paper is to answer an interesting research question about a causal relationship, and provide evidence suggesting the mechanism behind the relationship. The choice of method (difference-in-differences, regression discontinuity, or instrumental variables) depends on the nature of the quasi-experiment. The ‘etiquette’ is a way of routinizing the broader goal: To lay out an argument that convinces the researcher and the reader that the observed relationship is causal.

Before summarizing, it is important to state the limitations of causation-focused research in general and the quasi-experimental approach in particular. First, it is impossible to prove the validity of a quasi-experiment, such as whether one set of U.S. states serves as a legitimate control group for another or whether the exclusion restriction holds in instrumental variables. The credibility of any quasi-experimental work therefore relies on the plausibility of the argument for causality rather than on any formal statistical test. Second, external validity depends on the match between the treatments driven by the quasi-experimental variation and the overall sample needed to answer the research question. Quasi-experiments often require a focus on a narrow slice of the data and therefore it is important to consider the degree to which the results apply to a broader population. Third, all of these methods implicitly rely on throwing out variation in the data (the non-exogenous variation). In other words, they involve losing power in order to address exogeneity. This means that quasi-experimental work cannot use the R-squared as a useful summary of the appropriateness of the model. While R-squared or a comparison of log-likelihoods is very useful in many other contexts, it is not the focus of quasi-experimental papers.

We summarize by providing a list of questions to ask when reading and writing quasi-experimental papers:

1. Research question: Do we care if x causes y ?
 - Can x realistically be shifted in practice?
2. Identification strategy: Does x really cause y to shift?
 - Is the data structure understood?
 - Is there Model Free evidence?
 - Does the paper follow Difference-in-Differences, Regression Discontinuity, and Instrumental Variables etiquette?
 - Are there plenty of caveats and apologies that clarify the limitations of the analysis?
3. Mechanism: Why does x cause y to shift?
 - Is there evidence showing x causes y for the people that theory would predict?

References

- Almond, D., J. Doyle., A. Kowalski, and H. Williams (2010, May). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics* 125(2), 591–634.
- Altonji, J., T. Elder, and C. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113, 151–184.
- Anderson, E. T., N. M. Fong, D. I. Simester, and C. E. Tucker (2010). How sales taxes affect customer and firm behavior: the role of search on the internet. *Journal of Marketing Research* 47(2), 229–239.
- Anderson, S., R. Kellogg, A. Langer, and J. Sallee (2013). The intergenerational transmission of automobile brand preferences: Empirical evidence and implications for firm strategy. *Mimeo, University of Michigan*.
- Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton Press.
- Barreca, A. I., M. Guldi, J. M. Lindo, and G. R. Waddell (2011). Saving babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics* 126(4), 2117–2123.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119(1), 249–275.
- Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review* 102(6), 2472–2508.

- Busse, M., J. Silva-Risso, and F. Zettelmeyer (2006, September). \$1,000 cash back: The pass-through of auto manufacturer promotions. *American Economic Review* 96(4), 1253–1270.
- Cameron, C., J. Gelback, and D. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90(3), 414–427.
- Chevalier, J. and D. Mayzlin (2006). The effect of word of mouth online: Online book reviews. *Journal of Marketing Research* 43(345-354).
- Chintagunta, P. K. and H. S. Nair (2011). Discrete-choice models of consumer demand in marketing. *Marketing Science* 30(6), 977–996.
- Chiou, L. and C. Tucker (2012). Data Storage, Data Privacy and Search Engines. *Mimeo, MIT*.
- Chung, D. J., T. Steenburgh, and K. Sudhir (2014). Do bonuses enhance sales productivity? A dynamic structural analysis of bonus-based compensation plans. *Marketing Science* forthcoming.
- Dhar, T. and K. Baylis (2011). Fast-food consumption and the ban on advertising targeting children: the Quebec experience. *Journal of Marketing Research* 48(5), 799–813.
- Donald, S. and K. Lang (2007). Inference with difference in differences and other panel data. *Review of Economics and Statistics* 89(2), 221–233.
- Einav, L., A. Finkelstein, and M. Cullen (2010). Estimating welfare in insurance markets using variation in prices. *Quarterly Journal of Economics* 125(3), 877–921.
- Ellickson, P. B. and S. Misra (2011). Estimating discrete games. *Marketing Science* 30(6), 997–1010.

- Gelman (2010). *Experimental Reasoning in Social Science in Field Experiments and their Critics*. Yale University Press.
- Goldfarb, A. and C. E. Tucker (2011). Privacy regulation and online advertising. *Management Science* 57(1), 57–71.
- Gordon, B. R., R. Thomadsen, E. T. Bradlow, J.-P. Dube, and R. Staelin (2011). Revisiting the workshop on quantitative marketing and structural econometrics. *Marketing Science* 30(6), 945–949.
- Hartmann, W., H. Nair, and S. Narayanan (2011). Identifying causal marketing-mix effects using a regression discontinuity design. *Marketing Science* 30(6), 1079–1097.
- Heckman, J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *Quarterly Journal of Economics* 115, 47–97.
- Imbens, G. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142, 615–635.
- Imbens, G. and J. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47(1), 5–86.
- Lambrecht, A., K. Seim, and C. Tucker (2011). Stuck in the adoption funnel: The effect of interruptions in the adoption process on usage. *Marketing Science* 30(2), 355–367.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* 95, 391–413.
- Lee, J. Y. and D. R. Bell (2014). Why neighborhood social capital enhances social learning for experience attributes of products: Evidence from internet fashion retailing. *Marketing Science*.

- List, J. A. (2011, Summer). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives* 25(3), 3–16.
- Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.
- Mayzlin, D., Y. Dover, and J. Chevalier (2012). Promotional reviews: An empirical investigation of online review manipulation. *Working paper, University of Southern California*.
- Mela, C. F. (2011). Data selection and procurement. *Marketing Science* 30(6), 965–976.
- Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 12, 151–162.
- Qian, Y. (2008). Impacts of entry by counterfeiters. *Quarterly Journal of Economics* 123, 1577–1609.
- Qian, Y. (2014). Counterfeiters: Foes or friends? How do counterfeits affect different product quality tiers. *Management Science*.
- Reiss, P. C. (2011). Descriptive, structural, and experimental empirical methods in marketing research. *Marketing Science* 30(6), 950–964.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association* 100(469), 322–331.
- Shin, J., K. Sudhir, and D.-H. Yoon (2012). When to fire customers: Customer cost-based pricing. *Management Science* 58(5), 932–947.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science* 25(3), 289–310.
- Shriver, S. K., H. Nair, and R. Hofstetter (2013). Social ties and user generated content: Evidence from an online social network. *Management Science* 59(6), 1425–1443.

- Smith, J. and P. Todd (2008). Does matching overcome LaLonde’s critique of nonexperimental estimators. *Journal of Econometrics* 125(1-2), 305–353.
- Stock, J., J. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, 518–529.
- Sun, M. and F. Zhu (2013). Ad revenue and content commercialization: Evidence from blogs. *Management Science* 59(10), 2314–2331.
- Todd, P. (2008). *New Palgrave Dictionary of Economics*, Chapter Matching Estimators. Palgrave MacMillan.
- Tucker, C., J. Zhang, and T. Zhu (2013). Days on market and home sales. *RAND Journal of Economics* 44(2), 337–360.
- Wooldridge (2000). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.
- Zentner, A., M. Smith, and C. Kaya (2013). How video rental patterns change as consumers move online. *Management Science* 59(11), 2622–2634.
- Zhang, J. (2010). The sound of silence: Observational learning in the U.S. kidney market. *Marketing Science* 29(2), 315–335.