

MIT Open Access Articles

Conducting Research in Marketing with Quasi-Experiments

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Goldfarb, Avi, Tucker, Catherine and Wang, Yanwen. 2022. "Conducting Research in Marketing with Quasi-Experiments." *Journal of Marketing*, 86 (3).

As Published: 10.1177/00222429221082977

Publisher: SAGE Publications

Persistent URL: <https://hdl.handle.net/1721.1/144251>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution NonCommercial License 4.0



Conducting Research in Marketing with Quasi-Experiments

Avi Goldfarb, Catherine Tucker , and Yanwen Wang

Journal of Marketing
 2022, Vol. 86(3) 1-20
 © The Author(s) 2022



Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/00222429221082977
journals.sagepub.com/home/jmx



Abstract

This article aims to broaden the understanding of quasi-experimental methods among marketing scholars and those who read their work by describing the underlying logic and set of actions that make their work convincing. The purpose of quasi-experimental methods is, in the absence of experimental variation, to determine the presence of a causal relationship. First, the authors explore how to identify settings and data where it is interesting to understand whether an action causally affects a marketing outcome. Second, they outline how to structure an empirical strategy to identify a causal empirical relationship. The article details the application of various methods to identify how an action affects an outcome in marketing, including difference-in-differences, regression discontinuity, instrumental variables, propensity score matching, synthetic control, and selection bias correction. The authors emphasize the importance of clearly communicating the identifying assumptions underlying the assertion of causality. Last, they explain how exploring the behavioral mechanism—whether individual, organizational, or market level—can actually reinforce arguments of causality.

Keywords

quasi-experiments, marketing methods, econometrics

Online supplement: <https://doi.org/10.1177/00222429221082977>

Quasi-experimental methods have been widely applied in marketing to explain changes in consumer behavior, firm behavior, and market-level outcomes. “Quasi-experiment” refers to the use of an experimental mode of analysis and interpretation to data sets where the data-generating process is not itself intentionally experimental (Campbell 1965). Instead, quasi-experimental research uses variation that occurs without experimental intervention but is nonetheless exogenous to the particular research setting. Work using quasi-experiments in marketing settings has used events such as weather, geographic boundaries, contract changes, shifts in firm policy, individual-level life changes, and regulatory changes to approximate a real experiment. In each case, an external shock creates a source of exogenous variation that the researcher uses to establish a causal relationship between the variation and the outcome of interest.

Companies also use quasi-experimental methods to understand the consequences of key business actions. For example, Blake, Nosko, and Tadelis (2015) analyzed a quasi-experiment where eBay shut all the paid search advertising on Bing during a dispute with Microsoft but lost little traffic. These quasi-experimental results inspired a follow-up field experiment where eBay randomized suspension of its branded paid search advertising and found results consistent with the quasi-experiment. Reflecting the

importance of such methods at firms, some companies provide causal inference training for their data scientists (Crayton 2020; Rebecq 2020). The ability to make causal claims is highly valuable in academia and in practice. This article aims to help both marketing scholars and practitioners conduct and evaluate the credibility of quasi-experimental studies.

Quasi-experimental research, as in much work in applied statistics, begins with the equation $y = f(X, \varepsilon; \beta)$. The focus is then on whether a change in a single covariate x in the vector of X can be demonstrated to cause a change in y . This focus often enables the exploration of foundational questions in marketing, because marketers often have data representing the actions of many individual consumers or clients and need to

Avi Goldfarb is Professor of Marketing, Rotman Chair in Artificial Intelligence and Healthcare, Rotman School of Management, University of Toronto, Canada; and Research Associate, National Bureau of Economic Research, USA (email: agoldfarb@rotman.utoronto.ca). Catherine Tucker is Sloan Distinguished Professor of Management Science and Professor of Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology; and Research Associate, National Bureau of Economic Research, USA (email: cetucker@mit.edu). Yanwen Wang is Associate Professor of Marketing, Marketing and Behavioral Science Division, Sauder School of Business, University of British Columbia, Canada (email: yanwen.wang@sauder.ubc.ca).

understand the causal relationship between a particular x and y to make decisions about whether and how much x to use.

A marketing article that successfully uses the quasi-experimental econometric approach considers the following nine topics, which are echoed in the structure of this article:

1. Research Question: Do We Care Whether x Causes y ?
2. Data Question: How Can Researchers Find Data with Quasi-Experimental Variation in x ?
3. Identification Strategy: Does x Cause y to Change?
4. Empirical Analysis: How Can Researchers Estimate the Effect of x on y ?
5. Challenges to Research Design: What if Variation in x Is Not Exogenous?
6. Robustness: How Robust Is the Effect of x on y ?
7. Mechanism: Why Does x Cause y to Change?
8. External Validity: How Generalizable Is the Effect of x on y ?
9. Apologies: What Remains Unproven and What Are the Caveats?

We start by explaining why quasi-experimental scholars may appear obsessed with identification, and how this influences the choice of research question and data setting. Quasi-experiments come in different shades, ranging from an almost completely random exogenous shock to where the treatment assignment is only partly random. We suggest different frameworks to accommodate various levels of evidence depending on the strength of the underlying identification argument. We then turn to the importance of understanding the underlying mechanism behind the causal result. Typically, this means showing that the effect is largest where theory would predict and is smallest where theory would predict a negligible effect. We also emphasize that researchers need to be clear about the external validity of their study and apologize for what remains unconvincing.

Why the Focus on Identification?

Why are quasi-experimental scholars seemingly obsessed with identification? Identification is defined by the challenge that “many different theoretical models and hence many different causal interpretations may be consistent with the same data” (Heckman 2000, p. 47). However, effective decision making requires an understanding whether a measured relationship is indeed causal.

One way to describe this issue is through the “potential outcomes approach” developed by Jerzy Neyman, Donald Rubin, and others (Rubin 2005).¹ This approach starts with the insight that for any discrete treatment—which could be an event or explicit policy (D)—each individual i has two possible outcomes:

- y_{i1} if the individual i experiences the treatment $D_i = 1$, and
- y_{i0} if the individual i does not experience the treatment $D_i = 0$

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 at the same time. Therefore, only one outcome is observed for each individual at any point in time. 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 D , and those who do not, are different in unobserved ways.

Random assignment solves the inference problem, as the “unobserved ways” should not matter *ex ante* (Cook and Campbell 1979). Shadish, Cook, and Campbell (2002, p. 13) explain that “if implemented correctly, random assignment creates two or more groups of units that are probabilistically similar to each other on the average.” With enough people assigned randomly to one group or another, the only meaningful difference between the groups will be a result of the treatment.

Therefore, random assignment is often called the gold standard of identification (List 2011, p. 8). Angrist and Pischke (2009, p. 11) emphasize that “the most credible and influential research designs use random assignment.” That said, we should be clear that field experiments are merely a gold standard for being able to plausibly claim causality, not the gold standard for empirical work (Deaton 2009). Indeed, in many marketing situations, experiments are not feasible, appropriate, or affordable (Gelman 2010).

Quasi-experimental work, by contrast, is aimed to identify exogenous shocks or events that can approximate random assignment. Given that assignment is not random, a 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 and thereby mimic random assignment as closely as possible.

Research Question: Do We Care Whether x Causes y ?

The first and hardest stage in this process is identifying a question in which marketing scholars, managers, or policy makers actually care whether x causes y . This is difficult because many of the y s and x s for which we can measure a causal relationship are (unfortunately) uninteresting. Therefore, researchers who do quasi-experimental research do best if they start not with the data or an exogenous shock but instead start by asking themselves, “Suppose I convincingly showed that an increase in x increases y —who would care about this substantive issue?”

This means that the first stage requires the identification of a causal relationship that would be of interest to marketers or policy makers because their decisions will be usefully informed by a clear understanding of the consequences of a particular

¹ In describing these tools and their motivation, we build on numerous books and articles that have covered similar material for economics, policy, and sociology audiences (Angrist and Pischke 2009; Cook and Campbell 1979; Imbens and Wooldridge 2009; Meyer 1995).

action. As marketing technology and practices change, the number of measurable, interesting, and unanswered questions grows. A variety of editorials in this journal and elsewhere focus on how researchers can identify important issues. For example, the January 2021 special issue of the *Journal of Marketing* was dedicated to finding important marketing research questions, as highlighted in the editorial (Deighton, Mela, and Moorman 2021). Other editorials that discuss ways to identify important research questions are Van Heerde et al. (2021) and Chandy et al. (2021) in the *Journal of Marketing*, Schmitt et al. (2021) in the *Journal of Consumer Research*, Toubia (2022) in *Marketing Science*, and Grewal (2017) in the *Journal of Marketing Research*.

Data Question: How Can Researchers Find Data with Quasi-Experimental Variation in x?

Angrist and Pischke (2009, p. 7) explain that an identification strategy “describe[s] the manner in which a researcher uses observational data or data that is not generated as part of an intentional experiment, to approximate a real experiment.” They suggest first thinking of an ideal randomized experiment that can address the research question. This helps the researcher see clearly why an effect may not be identified causally in a non-experimental setting.

As Meyer (1995, p. 151) discusses, “Good natural experiments are studies in which there is a transparent exogenous source of variation in the explanatory variables that determine treatment assignment.” Unfortunately, there is no universally accepted interpretation of what it means to have a transparent exogenous source of variation. Therefore, Meyer (p. 151) emphasizes the importance of clarifying identification assumptions and understanding the institutional setting, stating, “If one cannot experimentally control the variation one is using, one should understand its source.” In the marketing context, Rossi (2014) discusses the dangers of using methods in which the source of the exogenous variation is either poorly understood or only weakly related to the correlation of interest.

Much of the work using quasi-experimental variation in marketing settings uses mundane but easily understood events such as contract changes, regulation, individual-level life changes, or shifts in firm policy that did not occur because of an anticipated effect on the outcome of interest. In some sense, some of the best sources of exogenous variation are mundane: nonmundane sources of variation such as global pandemics or earthquakes tend to be associated with other things happening that make it difficult to establish a clean causal relationship.

Table 1 lists several example quasi-experimental papers published in 2018, 2019, and 2020 in the *Journal of Marketing*, the *Journal of Marketing Research*, and *Marketing Science*. This table also summarizes the source of variation these articles use, spanning contractual changes; ecological variation (e.g., weather); geography; and macroeconomic, individual, organizational, and regulatory changes. It is useful to consider in turn why each of these sources of variation can approximate random assignment.

Contractual

To find plausibly exogenous variation in timing, it often depends on an argument that the exact timing of a measure is plausibly exogenous. Chiou and Tucker (2012) argued that the timing of a dispute between the Associated Press and Google was essentially random as it was influenced by a contract negotiated many years previously, and so the timing could be used to study the effect of the removal of content from news aggregators on downstream news websites.

Ecological

Generally, within-season variation in weather is plausibly exogenous. For example, Thomas (2020) uses quasi-experimental variation in actual and expected pollen counts. Key to the identification strategy is the focus on deviations from what was expected by firms.

Geographical

Work using geographical boundaries often exploits the fact that people who live on either side of a demarcated geographic border are similar enough to be thought of as being randomized across them. For example, by looking at a remote border of Maryland that was geographically isolated from the rest of the state, Anderson et al. (2010) were able to argue that the imposition of sales tax for those who lived on one side of the border was random, relative to those people who lived nearby but just happened to be over the state border.

Macroeconomic

It is also possible to take leverage of macroeconomic shocks. For example, Dubé, Hitsch, and Rossi (2018) use the Great Recession as a key source of the variation on household incomes over time. They exploit the within-household variation in private label shares associated with within-household changes in income and wealth. The identifying assumption is that, conditional on all other factors, including an overall trend, within-household changes in income and wealth are as good as randomly assigned or exogenous changes.

Individual

Plausibly exogenous variation can also be argued to occur at the individual level. For example, Bronnenberg, Dubé, and Gentzkow (2012) use consumer migration to new locations as a quasi-experiment to study the causal impact of past experiences on current purchases. They argue that while migration is not necessarily random, the precise direction of migration can be, at least with respect to local brand market shares.

Table 1. Examples of Quasi-Experiment Studies in *Journal of Marketing*, *Journal of Marketing Research*, and *Marketing Science* in 2018–2020.

Quasi-Experimental Variation			
General Category	Source	Article	Research Question
Contractual	Timing of American–Orbitz disputes to evaluate the absence of a major airline from a popular aggregator on consumer search	Akca and Rao (2020)	Who has more market power in the airline-aggregator relationships?
	Timing of the introduction of the <i>New York Times</i> paywall	Pattabhiramaiah, Sriram, and Manchanda (2019)	How does a paywall affect readership and site traffic?
Ecological	Variation in the forecast error of the pollen levels	Thomas (2020)	How much does advertising affect purchases of allergy products?
Geographical	Discontinuities in the level of advertising at the borders of DMAs	Shapiro (2020)	Does advertising affect consumer choice of health insurance?
	Discontinuities in the level of political ads at the borders of DMAs	Wang, Lewis, and Schweidel (2018)	How does political advertising source and message tone affect vote shares and turnout rates in 2010 and 2012 Senatorial elections?
Individual	Timing of users' adoption of a music streaming service	Datta, Knox, and Bronnenberg (2018)	How does a streaming service affect total music consumption?
	Variation in national ad exposures due to the local game outcomes	Hartmann and Klapper (2018)	How do Super Bowl ads affect brand purchases?
Macroeconomic	Variation in income and wealth due to the recession between 2006 and 2009	Dube, Hitsch, and Rossi (2018)	Do income and wealth affect demand for private label products?
Organizational	Discontinuity in the rounding rule that TripAdvisor uses to convert average ratings into displayed ratings	Hollenbeck, Moorthy, and Proserpio (2019)	How do online reviews affect advertising spending in the hotel industry?
	Timing of data breach and variation whether customer information was breached in a data breach event	Janakiraman, Lim, and Rishika (2018)	How does a data breach announcement affect customer spending and channel migration?
	Variation in timing of adoption of front-of-package nutritional labels across categories	Lim et al. (2020)	Do front-of-package nutritional labels affect nutritional quality for other brands in a category?
Regulatory	Timing of the Massachusetts open payment law	Guo, Sriram, and Manchanda (2020)	Do payment disclosure laws affect physician prescription behavior?
	Enforcement of minimum advertisement price policies	Israeli (2018)	What is the effect on violations if firms improve digital monitoring and enforcement of minimum advertised price policies?
	Timing of India's foreign direct investment liberalization reform in 1991	Ramani and Srinivasan (2019)	How do firms respond to foreign direct investment liberalization?

Notes: DMA = designated market area.

Organizational

Shifts in firm policy and organizational events can also be leveraged as a source of variation. For example, Janakiraman, Lim, and Rishika (2018) assess the change in customer behaviors between those whose information is breached and those whose information is not. The identification assumption is that the assignment of customers into the data breach group is likely to be random.

Regulatory

Many papers also use the timing of regulatory changes as a source of variation. The argument here is typically that though the imposition of regulation may not be random, the timing of the regulation is. For example, Tucker, Zhang, and Zhu (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, and Moorman, Ferraro, and Huber (2012) use a change in the standardized nutrition labels on food products required by the Nutrition Labeling and Education Act and investigate how the Act changes brand nutritional quality.

This discussion emphasizes that there are many potential sources of exogenous variation that can approximate a randomized experiment. We emphasize that typically the best papers focus on the research question first, and then imagine what the idealized experiment would look like to identify an actual quasi-experiment.

Identification Strategy: Does x Really Cause y to Change?

To convince a reader that an identification strategy is valid requires two steps. First, the researcher must explain where

the variation they are calling exogenous comes from. This requires institutional knowledge and careful research into the setting. Second, the researcher needs to demonstrate that the relationship between the variation and the outcome of interest is very likely driven by the relationship between x and y and not by some other factor.

To achieve the second requirement, it is useful to think about defending the experiment in terms of the exclusion restriction. Although the term “exclusion restriction” is often used specifically for instrumental variables, it is also a useful concept for other quasi-experiments. The exclusion restriction states that the quasi-experiment only affects y because it affects x .

There are a variety of ways in which the exclusion restriction can fail, and so researchers look for exogenous variation in x that will have no direct effect on y . For example, Shriver, Nair, and Hofstetter (2013) use wind speed as a quasi-experiment to 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 plausible way 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 windier surf locations. To address 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, Seim, and Tucker (2011), which examines the effect of delays in the early part of a banking technology adoption process on ultimate usage. Through a quasi-experiment 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 from others in terms of their laziness or motivation. To counter this challenge, the researchers present evidence that users are not different along any observable dimension.

The exclusion restriction can also fail because of spillovers between groups that receive the exogenous shock or treatment and those that do not. The assumption that treatment of unit i affects only the outcome of unit i is called the stable unit treatment value assumption (SUTVA) in the treatment literature (Angrist, Imbens, and Rubin 1996; Imbens and Rubin 2015). This is not a trivial assumption. For example, Akca and Rao (2020) use the 2011 Orbitz–American Airlines disputes as an exogenous event that led to a five-month period in which

American fares were not displayed on Orbitz. The authors use this dispute to identify which company was hurt the most in terms of site visits and purchases. The SUTVA requires a valid control group such that the Orbitz–American Airlines disputes have no spillover on that group. As a result, the authors chose not to use airfare- or hotel-booking websites as a control due to the possible spillovers from Orbitz to other websites where customers can purchase. Instead, the authors used consumers’ search of Lonely Planet as the control, because Lonely Planet is a travel website that is rarely used for bookings. The underlying idea is that an exclusion restriction cannot hold if the fact that one group was treated may also affect the control group’s behavior. The SUTVA is therefore part of an argument that researchers make about an appropriate exclusion restriction.

Importantly, there is no formula for a convincing explanation and defense of the empirical identification strategy in quasi-experiments. 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. 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.

Empirical Analysis: How Can Researchers Estimate the Effect of x on y ?

After establishing the identification assumption through the underlying framework of an exclusion restriction, the next step is to explore the data and conduct analysis that allows measurement of the effect of interest. This measured causal relationship is what has the potential to inform decision making. We discuss three different regression analysis frameworks using quasi-experiments: difference-in-differences (DID), regression discontinuity, and instrumental variables (IV). At the heart of all these strategies is a similar argument about the validity of the quasi-experiment.

Table 2 outlines eight key steps in the three regression analysis frameworks. As pointed out by Hahn, Todd, and Van der Klaauw (2001) and others, the techniques are very similar in terms of the underlying econometric theory. However, though similar in the conceptual ideas, in terms of practical implementation, presentation, and how the researcher should best reassure their audience about the validity of the technique, there are some differences, which we expand on. The three frameworks differ in the first four implementation steps. We discuss the first four steps for each of the three regression analysis frameworks and highlight the issues in common across the three analysis frameworks in the last four steps. We also emphasize that

Table 2. Quasi-Experimental Regression Analysis Frameworks.

	Difference-in-Differences	Regression Discontinuity	Instrumental Variables
Identification	Clarify the source of the shock, provide evidence why the shock can be seen as quasi-experimental, be clear on the identifying assumptions, and be transparent on the potential confoundedness.	Justify the source of the fixed threshold, and whether the assignment to the treatment is determined, either completely or partly, by the value of the predictor on either side of a fixed threshold.	Justify why the IV moves the endogenous covariate as if they are an experiment; explain the exclusion restriction.
Raw data	Test whether those who receive the treatment are similar to those who do not; whether the parallel assumptions are satisfied; illustrate the trajectory.	Provide evidence that the threshold is arbitrarily determined and not linked to underlying discontinuities in effects.	Regress the outcome directly on the instrument and show that the instrument has the expected direct effect.
Data analysis	Apply difference-in-differences regression framework in Equations 1 and 2 and adapt accordingly for other variations.	Apply regression continuity framework in Equation 3.	Report the first stage and determine whether the instruments are strong. Apply 2SLS in Equations 4 and 5 and conduct relevant tests.
Standard errors	Cluster at the level of treatment to account for within-unit correlation of the error term over time.	Use robust standard errors, do not cluster on a discrete variable	Cluster at the level of treatment to account for within-unit correlation of the error term over time.
Robustness checks		Conduct multiple robustness checks.	
Mechanism checks		Measure mediator variables or show moderation analysis.	
External validity		Discuss the assumptions required to capture the ATE.	
Apologies and caveats		Apologize for all that is still unproven and give caveats.	

many excellent papers do not implement each step, and this description is not intended to lead to unproductive dogmatism.

All of these methods implicitly rely on throwing out variation in the data that is not exogenous. In other words, they involve losing power to support the exogeneity assumption. This means that quasi-experimental work cannot use the R-squared as a useful summary of the appropriateness of the model. Ebbes, Papies, and Van Heerde (2011) provide some useful evidence. While R-squared or a comparison of log-likelihoods is very useful in many other contexts (e.g., forecasts), benchmarking quasi-experimental analyses against other methods by using the R-squared will be misleading.

Difference-in-Differences Analysis

A standard DID analysis compares a treatment group and a (quasi-) control group before and after the time of the treatment. The “treatment” is not truly a random experiment but, rather, some “shock.” Unlike a simple comparison (or single-difference) analysis, DID methods generate a baseline for comparison between the treatment and the control group. By highlighting the change in the treatment group relative to the control group, DID enables the researcher to control for many of the most obvious sources of heterogeneity across groups.

Goldfarb and Tucker (2011a) is an example of a DID paper. The authors 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. The paper uses data on the success of nearly 10,000 online display advertising campaigns in Europe, the United States, and elsewhere between 2001 and 2008. The authors compare the change in effectiveness of the ad campaigns inside and outside Europe. Therefore, the first difference is the change in the campaign effectiveness, and the second difference is the change in Europe relative to elsewhere. Compared with before the regulation, ad campaigns became 2.8% less effective in Europe after the regulation. In contrast, compared with before the European regulation, ad campaigns became .1% more effective outside of Europe after the European regulation was implemented.

Identification of Difference-in-Differences

The first step is to clearly lay out the identifying assumptions. Goldfarb and Tucker (2011a, p. 63) state that “the identification is based on the assumption that coinciding with the enactment of privacy laws, there was no systematic change in advertising effectiveness independent of the law” and that “the European campaigns and the European respondents do not systematically change over time for reasons other than the regulations.” A substantial portion of the paper is devoted to providing empirical evidence regarding whether (1) European ad agencies invest

less in their ad creatives relative to non-European ad agencies after the laws, (2) the demographic profile of the respondents is representative of the general population of internet users, and (3) there may have been a change in European consumer attitudes and responsiveness to online advertising separate from the Privacy Directive.

The analysis of consumer attitudes and ad responsiveness is based on a concern about unobservables, specifically whether there are alternative explanations for the measured changes in the attitudes of survey participants toward online advertising that were separate but contemporaneous with the change in European privacy laws. To check for such unobserved heterogeneity, Goldfarb and Tucker (2011a) examine the behavior of Europeans on non-European websites that are not covered by the European Privacy Directive to see if a similar shift in behavior can be observed, and they find evidence that changes in behavior are connected with the websites covered by the law, rather than the people taking the survey. The identification exclusion criterion is further validated by a mirror image of the falsification test by looking at residents of non-European Union (EU) countries who visited EU websites. When residents of non-EU countries visit EU websites, the ads are less effective in the postperiod. In contrast, when residents of these non-EU countries visit non-EU websites, there is no change in effectiveness before and after the EU regulation. Therefore, the results appear to be driven by what happens at EU websites rather than by a difference in how Europeans behave relative to non-Europeans.

Raw Data Exploration of Difference-in-Differences

The second step is to explore the raw data. Before applying the DID framework, it is important to explore the raw data to assess whether the quasi-experiment appeared to have an effect. For example, when a treatment occurs in the middle of a time series, 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 diverged from the control group.

Researchers should also assess whether their quasi-experimental setting meets the parallel trend assumption while exploring their raw data. This involves demonstrating that behaviors were similar in the period prior to the policy change across the treatment and control groups. Depending on the length of the time period, this can be done by conducting two-sample mean comparisons for each pre-treatment period or by running a linear regression and looking at the time trend differences between the control and treatment groups. It is also often ideal to simply plot the raw data to support this point.

Though it is desirable and convincing if the main effect of interest can be seen through descriptive statistics or visualization, we caution that this is not always possible. This may happen because effect sizes are small—as they often are in advertising—or because there is variation in the data that is best addressed using a regression framework.

Analysis of Difference-in-Differences

Although a DID regression can be represented in a 2×2 table, it is usually analyzed with regression analysis to allow researchers to control for factors that may change over time and across individuals. The simplest version of this regression is as follows:

$$y_{it} = \alpha_1 \text{Treatment Group}_i + \alpha_2 \text{After Treatment}_t + \beta \text{Treatment Group}_i \times \text{After Treatment}_t + \gamma X_{it} + \varepsilon_{it}, \quad (1)$$

where y is the outcome of interest; i represents the individual, firm, or other cross-sectional unit of interest; t represents the time period; and ε_{it} represents the error. The key focus of the DID specification is on β , which captures the explanatory power 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 the treatment and control groups in the same way.

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 set 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 DID regression is

$$y_{it} = \beta \text{Treatment Group}_i \times \text{After Treatment}_t + \gamma X_{it} + \mu_i + \tau_t + \varepsilon_{it}, \quad (2)$$

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 Treatment Group_i and After Treatment_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.²

Though changes over time are common, DID methods do not require a time-series component. For example, Goldfarb and Tucker (2011b) examine the impact of offline advertising restrictions on prices for keyword advertisements. The first difference is the keyword ad prices in states that have restrictions compared with states that do not. The second difference is the keywords that are affected by the restrictions compared with the keywords that are not.

For quasi-experimental analyses that do examine changes over time, another tweak is that quasi-experimental treatment can occur at different times, meaning that individuals are treated at different times and that the *After Treatment* variable

² For 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.

can change with subscripts i and t . For example, Chevalier and Mayzlin (2006) study how a book review posted on Amazon affects sales of that book on Amazon, compared with 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 After Treatment period occurs at different times for different books. Athey and Imbens (2022), Borusyak, Hull, and Jaravel (2022), Callaway and Sant'Anna (2020), De Chaisemartin and d'Haultfoeuille (2020), Goodman-Bacon (2021), and Roth et al. (2022) explore the effects of variation in treatment timing. The issue is that because a fixed-effects DID estimator is a weighted sum of the treatment effect in each group and at each period, even though the weights sum to one, negative weights may arise when there is a substantial amount of heterogeneity in the treatment effects over time. A related concern has been highlighted by Gibbons, Serrato, and Urbancic (2017), who emphasize the problems that occur when both the treatment effect and treatment variance vary across groups.

This means that researchers should be cautious in summarizing time-varying treatment effects with a homogeneous treatment effect as in the two-way DID framework if there is a substantial timing dimension. To address these issues, researchers have proposed a variety of estimators that allow for a cleaner comparison between the treated group and the control group. Both Callaway and Sant'Anna (2020) and De Chaisemartin and d'Haultfoeuille (2020) propose new estimands to estimate treatment effects in the presence of heterogeneity across groups and over time.³ Another approach is taken by Sun and Abraham (2021), who discuss corrections that should be applicable in a situation where leads or lags might be expected.

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

Regression Discontinuity Analysis

Regression discontinuity is a quasi-experimental technique in which the "experiment" relies on an exogenous arbitrary threshold. As Imbens and Lemieux (2008, p. 616) put it, "The basic idea behind the RD [regression discontinuity] 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."

Identification in regression discontinuity. Regression discontinuity may be particularly useful to marketing scholars. Hartmann, Nair, and Narayanan (2011) argue that many marketing

interventions are based on thresholds of real or expected consumer or firm behavior. For example, direct mail companies use the scoring policies for recency, frequency, monetary 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.

Similarly, government policies based on firm size can provide a useful identification strategy for marketing scholars. For example, requirements for firms to post calories, undertake layoffs, and provide benefits often 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 effect of the policies on firm behavior.

A regression discontinuity design implies that treatment is assigned depending on whether a continuous score z_i crosses a cutoff \bar{z} . The analysis then focuses on whether there is a change in the outcome of interest y in the neighborhood of \bar{z} (Hartmann, Nair, and Narayanan 2011). In general, if a threshold is used as the source of the quasi-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.

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 to get the incentive. In contrast, those with exactly \$50 in spending might have selectively chosen to spend exactly enough to get an incentive that they planned to use. It is important to address the potential for such concerns directly.

This is reflected in a debate in economics about the effect of thresholds for low birth weight on medical outcomes. In an initial study, Almond et al. (2010) used the fact that birth weight threshold of 1.5 kg is used to determine whether the newborn receives intensive medical treatment. In a critique of this work, Barreca et al. (2011) show that the children placed just at the cutoff seem to have significantly worse outcomes than babies on either side of the cutoff. This is evidence against use of this discontinuity for identification. Barreca et al. 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" (p. 2119).

Raw data exploration of regression discontinuity. Once the researcher has found a regression discontinuity setting, the first step is to explore whether the discontinuity is arbitrary and linked to discontinuities in any other variables. For example, Hollenbeck, Moorthy, and Proserpio (2019) examine the relationship between online reviews and advertising spending in the hotel

³ Detailed implementation steps are provided in `fuzzydidi` and `did_multipl` in STATA and the `did` package in R.

industry. They exploit the regression discontinuity design of the rounding rule that TripAdvisor uses to convert the average ratings of reviewers into the nearest half or full star (i.e., a rating of 3.74 is shown as 3.5 stars while a rating of 3.75 shown as 4 stars), building on work by Luca (2011). The key identification argument is that the rounding mechanism creates discrete, random variations in *perceived* quality around the rounding threshold and is independent of a hotel's *true* quality.

A threat to the arbitrary discontinuity threshold would be that hotels manipulate their average ratings around the rounding thresholds. Hollenbeck, Moorthy, and Proserpio (2019) argue that if there is upward manipulation of ratings, there would be relatively few firms with average ratings just below the thresholds and a clump of firms with average ratings just above the thresholds. They show instead that the density of average ratings is uniform, with neither bumps nor dips above or below the round thresholds. They provide additional empirical evidence that characteristics of the hotels do not differ systematically above or below the threshold. Neither do they observe discontinuities in other key variables such as hotel prices and the number of five-star reviews.

Analysis of regression discontinuity. The equation used for regression discontinuity can be written for panel data as

$$y_{it} = \beta I(z_{it} \geq \bar{z}) + \gamma X_{it} + \mu_i + \tau_t + \varepsilon_{it}. \quad (3)$$

Here β is the treatment effect, the parameter of interest. X_i represents covariates. $I(z_{it} \geq \bar{z})$ is an indicator function that equals one when $z_{it} \geq \bar{z}$ and zero otherwise. One final consideration is how to select the appropriate bandwidth for a regression discontinuity design, which is the question of how one decides on the sample to analyze, in terms of how far away the people in the sample are from the threshold where the discontinuity occurs. In general, such decisions have often been rather ad hoc, but there is an emerging literature that can help guide the researcher into thinking about how to take a more conservative approach to selecting bandwidth given the data at hand (Cattaneo, Titiunik, and Vazquez-Bare 2020). The researcher should also ensure that their results are not sensitive to the choice of bandwidth. As with other quasi-experimental methods, the validity of the method cannot be statistically proven. Therefore, substantial emphasis must be placed on the explanation and defense of the quasi-experiment using raw data.

Instrumental Variables Analysis

The quasi-experimental perspective on IVs is somewhat different from the standard treatment in econometrics textbooks, which focuses on simultaneous equations and a more structural approach. The differences relate to justification and interpretation. The quasi-experimental approach emphasizes that the shocks that move the instrument should behave as if they are an experiment. The quasi-experimental approach gives a sense of the sign, significance, and magnitude of the causal effects. The structural approach emphasizes that the shocks should be

motivated by an economic model that explains the exclusion restriction. The IV approach used in structural models gives elasticities that can be used to generate counterfactuals outside of the sample. Despite these differences in interpretation, it is important to remember that the underlying mathematics is identical.

Identification of instrumental variables. The basic idea behind using IVs is that the covariate of interest x contains both useful variation (to identify the causal effect of interest) and less useful variation (that confounds the effect). A good instrument z is strongly correlated with the useful variation but uncorrelated with the confounding variation. In other words, the researcher only uses the variation in x that can be explained by the exogenous shifter z .

The standard two-stage model involves two steps. In the first-stage regression, a fitted value of \hat{x}_i can be obtained by regressing x on instrument z and covariates W :

$$x_i = \gamma z_i + \vartheta W_i + \eta_i. \quad (4)$$

In the second-stage regression, the IV estimator $\hat{\beta}$ is obtained by regressing the outcome y on the fitted value of \hat{x} and covariates W :

$$y_i = \beta \hat{x}_i + \phi W_i + \varepsilon_i. \quad (5)$$

The identification of the effect of x on y relies on the following “reduced form.” Inserting the predicted x to the y equation will give Equation 6. Here, $\hat{\phi}$ is used to highlight that when regressing y directly on instrument z and covariates W , the estimated covariate coefficient is rescaled as $\hat{\phi} = \beta\vartheta + \phi$.

$$y_i = \beta \gamma z_i + \hat{\phi} W_i + \varepsilon_i. \quad (6)$$

Therefore, from the quasi-experimental point of view, an instrumental variable can be seen as a treatment that affects the endogenous covariate directly. 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. There are many ways of operationalizing instrumental variables, and this can be a place for highly technical tools. We emphasize the simplest two-stage least squares (2SLS) approach, but the intuition behind the role of instrumental variables as an identification strategy remains regardless of functional form assumptions. Using two stages enables 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 .

Returning to Shriver, Nair, and Hofstetter (2013), while the paper adds some additional necessary nuance to the estimation to fit the particular situation, the intuition on causality measures the impact of wind (the instrument z) on social ties (the outcome of interest y). This will be $\beta\gamma$. The relationship of interest, however, is the impact of posts (x) on social ties (y), which is measured as β .

IV can be a less transparent solution to identifying causal effects compared with the other two analysis frameworks discussed previously (for a detailed discussion, see Rossi [2014]). The distinction between the relationship of interest (β) and the direct estimate from the quasi-experiment ($\beta\gamma$) means that it is sometimes harder to visualize how the quasi-experimental variation works in IVs.

Transparent communication of IV analysis is difficult for three reasons. First, in contrast to the binary nature of the exogenous variation in DID and regression discontinuity, instruments are often continuous. This makes it more difficult to communicate the intuition for why the variation is exogenous to the potential for omitted variables or simultaneity. The ability to use continuous instruments (and multiple instruments) can also be seen as a strength of IV techniques. They enable 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.

Second, weak instruments are a challenge. Instrumental variables techniques are consistent but biased, and this bias can matter even in seemingly large samples (Stock, Wright, and Yogo 2002). Weak instruments can lead to incorrect inference in which the bias of the weak instrument dominates the potential bias of the omitted variables.

Knowing the context and the institutional setting can be invaluable in identifying strong IVs. For example, Moorman, Ferraro, and Huber (2012) derived their instruments for brand taste and price from the authors' intimate knowledge of the regulation and food industry. There are also recent advances in econometric methods that allow for more accurate presentation of statistical significance when instruments are weak (Lee et al. 2021). As Angrist and Kolesár (2021) point out, many of the challenges of weak instruments are magnified when authors use multiple instruments to deal with multiple sources of endogeneity. By contrast, a focus on a single endogenous variable with a single source of endogenous variation has attractive statistical properties as well as being more transparent to the reader.

Third, many researchers present IV results with different tests and with different norms. This makes it difficult to read and assess the validity of papers with instruments.

Raw data exploration and analysis of Instrumental Variables.

Angrist and Pischke (2009, pp. 212–13) provide a sequence of steps to follow in an attempt to standardize practice. In presenting this list, we hope that it does not lead to unproductive dogmatism, and we emphasize that this is just one possible way to communicate the rationale behind a causal interpretation of the results. Still, we hope that in following these steps to the extent possible, marketing scholars can avoid being subject to many of the criticisms highlighted by Rossi (2014). The steps are as follows:

1. Regress the outcome directly on the instrument. When using IV techniques, it is also desirable to show the reduced form result of regressing the outcome directly on the instrument. Because this is an ordinary least squares regression, it is unbiased. At the very least, the researcher should be confident that the instrument (z) has the expected direct effect on the outcome (y).
2. Report the first stage. Assess whether the signs and magnitudes of the coefficients make sense.
3. Report the F-statistic on the excluded instruments. This helps determine whether the instruments are weak. Stock, Wright, and Yogo (2002) advise that F-statistics below 10 in case of only one instrument suggest weak instruments, though, as Angrist and Pischke (2009, p. 213) note, "Obviously this cannot be a theorem." Similarly, Rossi (2014) suggests reporting the first stage with and without the instruments to document the incremental impact of the instruments on the R-squared.
4. 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. Presenting the results separately also helps the reader understand the intuition behind the quasi-experiment underlying each instrument—whether the multiple instruments use different variation in increasing the exogenous shift in x . If there are multiple instruments, an overidentification test such as the Sargan–Hansen J can be performed to test whether all instruments are uncorrelated with the 2SLS residuals.⁴ However, given the difficulty of identifying a robust instrument, it is unusual for researchers to have convincing cases for multiple instruments in a way that leads their regression to be overidentified. In other words, increasingly, standard practice is to focus on one instrument rather than many (Angrist and Kolesár 2021).
5. Conduct a Hausman test comparing ordinary least squares and instrumental variables. If the results change, reflect on whether they change in a direction that makes sense given the power of the instrument. Do not interpret the results of the Hausman test to prove that the endogeneity problem is irrelevant. As noted by Rossi (2014), the instrument may not be valid and therefore the test would be uninformative.
6. Assess whether there is a weak instrument problem. For example, in a linear model, compare the 2SLS results with the limited information maximum likelihood results. When there is a weak instrument, the two-stage least square estimators are biased in small sample. Limited information maximum likelihood estimators have better small sample properties than 2SLS with

⁴ For example, as of 2022, IV estimation can be produced by *ivreg2* in STATA. Under i.i.d. error assumption, the command *estatoverid* provides the Sargan test. When the estimation is done with generalized method of moments (i.e. *gmm2s* is specified in *ivreg2*), the test of overidentifying restrictions becomes the Hansen J statistic.

weak instruments. If the two estimates are different, there may be a weak instrument problem. Any inconsistency from a small violation of the exclusion restriction gets magnified by weak instruments.

Presentation of Results and Clustering of Errors

Regardless of which regression analysis framework to employ, presentation of baseline estimates and standard errors, along with a set of robustness checks (Van Heerde et al. 2021) is standard. 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. The economic magnitude of the coefficients should be discussed, both with respect to changes in the covariate of interest and relative to the range and standard deviation of the covariate and dependent variable.

A key issue in quasi-experimental 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 several 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; however, the choices of the same individual in many time periods are likely to be correlated. Bertrand, Duflo, and Mullainathan (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 downward. 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 and useful 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 such as General Mills, the brand such as Cheerios, or the sub-brand such as Honey Nut Cheerios? The answer depends on the data and research question. If the data are at a lower unit level (e.g., individuals) than a treatment that takes place at the firm level, cluster the standard errors at the level of the treatment. A useful perspective on this is provided by Abadie et al. (2017), who remind researchers that the major driver for clustering should be the experimental design rather than simple expectations of correlation. More recently, there has been evidence suggesting that it is undesirable to cluster on the variable that determines whether that observation is subject to the regression discontinuity design (e.g., age). The answer is often instead simply to reduce the bandwidth across which the regression discontinuity is studied (Kolesár and Rothe 2018).

Clustered standard errors rely on consistency arguments and large samples. With a small number of clusters, alternative

methods are needed, such as those developed by Cameron, Gelback, and Miller (2008), Conley and Taber (2011), and Hagemann (2019). For example, Elberg et al. (2019) investigate consumers' dynamic responses to price promotions in a retail setting that involved randomly assigning ten supermarkets into varying promotion depths. Given that treatment takes place at the store level while the observation is at the consumer level, each consumer's effective contribution to reducing standard error estimates is likely to be lower than in a setting where there is no correlation across observations. However, given the relatively small number of stores/clusters available in this setting, the authors implement the wild bootstrap procedure, as proposed by Cameron, Gelback, and Miller (2008), to correct for downward bias potentially induced in small samples. However, Canay, Santos, and Shaikh (2021) show that even this approach requires rather large assumptions.

Challenges to Research Design: What if Variation in x is not Exogenous?

A more general point is that quasi-experiments range in how plausible the exogenous variation underlying the paper is, ranging from cases where the allocation is almost completely random to less clear cases where a firm or consumer assignment to treatment or control is partly random and partly an endogenous choice. Perhaps the ideal thought experiment here is Zhang (2010), whose treatment and control were a pair of kidneys from the same person. Zhang finds that in the United States, even identical kidneys from the same donor are received differently depending on the observed number of rejections preceding the recipient in the queue. Most research settings are less favorable. In such settings, it is often useful to combine different approaches in the same paper. For example, Qian (2008) combines a DID strategy with counterfeit entry as the treatment with a convincing and high-powered instrument on government regulation.

Still, there will be situations where a compelling exclusion restriction is lacking or the treatment–control allocation appears far from random. If the treatment and control groups are substantially different in the pretreatment or if the treatment appears to be applied based on selected characteristics, the control group is unlikely to be a good proxy for the counterfactual, and the quasi-experiment may be less likely to be valid.

We provide a discussion of three methods that are further steps researchers can take when comparability between the control and treatment groups is violated. They vary in terms of the observed and potentially unobserved differences between the control and treatment groups. Table 3 provides a summary of the frameworks and when to apply them. The table emphasizes that researchers should be cautious about applying matching methods or correction for selection bias on the grounds that there are no plausible exclusion restrictions, because these methods still require the researcher to make an argument about an exclusion restriction. The technical details of matching methods or selection bias correction are different from the three methods described previously, but the idea is similar in

Table 3. Steps if Researchers Are Worried They Do Not Meet the Exclusion Restriction.

	Propensity Score Matching	Synthetic Control	Selection Bias Correction
Assumptions	Observable control variables are capable of identifying the selection into treatment and control conditions	The counterfactual outcome of the treatment units can be imputed in a linear combination of control units in the absence of treatment.	The unobservables that enter the treatment selection and the outcome are jointly distributed as bivariate normal.
Identification	The exclusion restriction can be met conditional on the variables in the match.	The exclusion restriction can be met conditional on the pretreatment outcomes.	There is at least one variable for which a compelling argument can be made for the exclusion restriction in the selection equation.
Settings	When matching is done to control the treatment and control pretreatment outcomes on a number of cross-sectional covariates.	When the focus is on the evolution of the outcome and the pretreatment time period has rich data on treatment and control groups.	When the allocation to the treatment condition is not fully random.
Caveats	Assess the degree of overlap after matching, and assess sensitivity to potential selection on unobservables. Still need to justify the exclusion restriction.	Harder to interpret the weights used to create the “synthetic control.” Still need to justify the exclusion restriction.	Justification of why certain observables only affect treatment selection but not the outcome variable. Still need to justify the exclusion restriction.

nature. The main goal is to bring in additional data to create control and treatment groups that are like those in quasi-experiment studies.

Propensity Score Matching

Matching methods, pioneered by Rosenbaum and Rubin (1983), have been developed such that the outcomes of the treated are contrasted only against the outcomes of comparable untreated units. Many published articles in marketing have used propensity score matching when comparability between the control and treatment groups is violated. An assumption of propensity score matching is that there are observable control variables capable of identifying the selection into treatment and control conditions. This is not a trivial assumption. It suggests that propensity score matching is only good if the exclusion restriction is met conditional on the variables in the match. Any matching procedure to make the control and treatment more similar in the observables can be seen as a flexible functional form with adding “control variables” to an analysis framework. Propensity score matching requires subject-matter knowledge regarding the role of covariates in the treatment assignment decision and whether the exclusion restriction is satisfied conditional on the covariates. Therefore, we caution against applying matching methods without convincing justification of exclusion restriction.

It is difficult to identify a standard procedure for propensity score matching. We refer to Imbens and Rubin (2015) as a good starting point. The general objective of propensity score matching is to estimate a score such that the distribution of all the observed variables and behaviors among the treated units is similar to that among the control units. In this discussion, we consider the set of treated units to be fixed a priori. Four steps are involved in the propensity-score-matching procedure.

First, choose a functional form of the propensity score. The basic strategy uses logistic regression to model the probability of receiving the treatment given a set of observables. Second,

measure the distance and apply a matching algorithm. Several possible matching methods are available including, for example, nearest-neighbor matching based on the distance in the estimated propensity score or multiple matching using all controls within some distance from the treated unit. Third, assess the degree of overlap in the distribution of the linearized propensity score after matching. Researchers typically plot and compare the histogram-based estimate of the distribution of the linearized propensity score (logarithm odds ratio) for the treatment and control groups. To inspect the match quality, it is useful to show tables on the distribution of the estimated propensity scores and the mean values of some key variables for the treated and untreated over different propensity score intervals.⁵ Fourth and finally, calculate the average treatment effect (ATE) with the matched sample using, for example, the DID regression analysis framework discussed previously.

There are at least two caveats regarding propensity score matching. First, the model for the propensity score may be misspecified. In that case, the balance in covariates conditional on the estimated propensity score may not hold, and the credibility of subsequent inferences may be compromised. This calls for a careful discussion on the role of covariates in the treatment assignment decision. Specifically, it is important to provide a discussion of whether the covariates can be considered exogenous to the treatment. Second, regardless of the number of observed covariates used, propensity score matching does not account for the potential selection on unobservables in treatment assignment. It is important to explain why controlling for observables will address concerns with the exclusion restriction or why unobservables are not an issue in treatment assignment.

⁵ The propensity score matching algorithm can be found in multiple statistical packages as of 2022, for example, the PSMATCH2 module in Stata.

Synthetic Control Methods

In some cases, even the closest match may not be close enough. This is particularly relevant when researchers are interested in how an event, regulatory intervention, or firm policy change affects the evolution of the outcome of interest, in contexts where only a modest number of treated units (possibly only a single one) and control units are observed for a large number of periods before and after the event. Two aspects make this setting different from the typical use of the propensity-score-matching method. First, matching is done over the pretreatment outcomes in each period rather than a number of covariates. Second, the number of control units and the number of pretreatment periods can be of similar magnitude. Synthetic controls use a different convex combination of the available control units (Abadie, Diamond, and Hainmueller 2010; Abadie, Diamond, and Hainmueller 2015; Doudchenko and Imbens 2016). The intuition behind this method is that the created synthetic control unit closely represents the treated unit in all the pretreatment periods and affords time-varying causal inference on the trajectory of the outcome of interest.

Synthetic control has been used in multiple recent studies with quasi-experimental design (Abadie 2021). For example, Guo, Sriram, and Manchanda (2020) analyze the causal effect of industry payment disclosure on physician prescription behavior, Wang, Wu, and Zhu (2019) assess the impact of mobile hailing technology adoption on drivers' hourly earnings, and Pattabhiramaiah, Sriram, and Manchanda (2019) study the causal effect of online paywalls on the sales revenues of newspapers.

Like propensity score matching, synthetic control methods are statistically rich, but they do not replace a carefully thought-out exclusion restriction and identification argument. Put differently, if propensity scores or synthetic controls appear to work when the treatment and control group are not similar, it is important to explain why controlling for observables will address issues with the exclusion restriction. In many cases, such explanations are weak and the exclusion restriction is unlikely to hold. Recent work in economics emphasizes this by showing the benefits of combining a synthetic control method with a strong exclusion restriction (Arkhangelsky et al. 2021).

Selection Bias Correction Method

Many papers written in marketing involve a comparison of potentially different groups that reflect endogenous choices by companies or consumers where the allocation to the treatment condition is not fully random. For example, Gill, Sridhar, and Grewal (2017) assess if the introduction of the free mobile app in a business-to-business context increases sales revenues from buyers who adopted the app. In an ideal setting, the company could randomize the treatment, then observe sales from buyers who did not get the app and sales from buyers who did get it. However, this company's app was available to all buyers. Therefore, the buyers' app adoption is not random, and self-selection into the treatment (adoption) group needs to be addressed. Omitted variables that drive strategic app adoption could correlate with the sales from these buyers.

When this happens, it is sometimes useful to estimate a Heckman selection model (Heckman 1978), which explicitly models selection into the treatment as a two-step process. As Wooldridge (2000, p. 564) pointed out, the exclusion criterion is still key to the identification of the treatment effect of interest in the two-step estimation procedure. Without the exclusion criterion, the effect of the treatment is identified only due to the nonlinearity in the functional form (specifically through the inverse Mills ratio). This may lead to severe collinearity and imprecision in the standard errors. More importantly, without a strong and credible exclusion restriction, identification in this setting is driven by the assumed functional form.

In other words, although the Heckman correction will provide an estimate without an exclusion restriction, that estimate depends entirely on the assumption that the error structure is bivariate normal. When there is an argument for the exclusion restriction, a selection model is helpful. In the absence of the exclusion restriction, even if combined with other techniques such as propensity score matching, the results would be identified off the functional form assumption alone. Put differently, if one of the covariates in the correction equation satisfies the exclusion restriction, then it is the variation in that variable that identifies the control for selection. In contrast, if the covariates in the first step are all also in the second step, then it is only the assumed error structure that identifies the control for selection.

There are both similarities and differences between selection bias correction and instrumental variable approaches. There are also similarities with the control function approach in terms of the importance of functional form assumptions on the errors in the absence of an exclusion restriction. Control functions are not part of the standard quasi-experimental toolkit, so we do not provide a detailed discussion. The selection bias correction approach uses the instrument to control for the effect of unobservables, while the instrumental variable approach attempts to eliminate the threat of endogeneity by only leveraging the useful variation created by the instrument. Yet, the two approaches share the basic idea of using an exclusion criterion (or instrument). Ultimately, both rely on the ability to find an exclusion restriction that creates useful and exogenous variation. This is why we emphasize the importance of identification in quasi-experiments and caution against blindly applying a correction for selection bias without carefully thinking about the identification assumption and providing a justification for why the exclusion restriction holds. Selection bias correction approaches are therefore only useful for causal inference in the presence of a strong credible exclusion restriction.

Robustness: How Robust is the Effect of x on y ?

The specific robustness checks chosen will depend on the exact context. With electronic appendices and increasingly cheap computation, it is possible to show robustness to a large number of alternative specifications. Here, empirical work with quasi-experimental methods differs substantially from research using forecasting models. The aim is not to show

one specification (or model) and defend it. Instead, the idea is to show that the sign, significance, and magnitude of the estimate of β remain broadly consistent across a vast range of possible models (Van Heerde et al. 2021). Often these robustness checks are dropped from the published version of the article, though they are very useful in the referee process and can end up as part of an online appendix. The following subsections describe some examples of useful robustness checks.

Different Controls

Compare the coefficient of interest in the models with and without controls. For example, if the coefficient changes from 2.5 to 3.5, then this change (+1.0 in this example) is informative about how big the impact of the omitted variables has to be relative to the observed controls for the omitted variables to drive the result. Altonji, Elder, and Taber (2005) provide a method 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 disappear. The method is based on Rosenbaum bounds (DiPrete and Gangl 2004; Rosenbaum 2002). It has been applied in the marketing literature by Manchanda, Packard, and Pattabhiramaiah (2015) and extended by Shin, Sudhir, and Yoon (2012). Although the formal method is useful, as discussed in Oster (2019), many researchers (Anderson et al. 2015; Mayzlin, Dover, and Chevalier 2014) use the more basic insight that there is information in the impact of the controls on the measured effect of interest. This does not mean that results are invalid if the controls do change the estimated effect substantially, but documenting that adding seemingly relevant controls does not change the results can provide further support for the causal interpretation.

Different Functional Forms

Results should not depend on arbitrary choices of functional form. For example, if using a linear probability model, show robustness to logit and probit. The choice between linear probability models and nonlinear 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.

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, researchers should be comfortable that the results are robust to the arbitrary choice of whether the period studied is 2002 to 2006, 2000 to 2008, 1995 to 2015, and so on.

Different Dependent Variables

There might be several different dependent variables that relate to the outcome of interest. Showing robustness to these related outcomes increases confidence in results.

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 (e.g., as measured by a propensity score). Researchers can also choose how to define the treatment group.

Placebo Tests

The idea of a placebo test is to repeat your analysis using a different part of the data set where no intervention occurred. For example, if the quasi-experimental shock happens this year, instead of comparing the difference in the outcome between last year and this year between the control and treatment groups, you can conduct a placebo test by redoing the analysis and compare the difference in the outcome between the control and treatment groups using periods with no intervention shocks. Alternatively, analysis can be conducted on an outcome that should be unrelated to the intervention being studied. The goal is to establish a null effect when there is not supposed to be one.

It is unlikely that every robustness check will yield the same level of significance or the same-sized point estimate as the initial specification. Researchers (and reviewers) should therefore not expect every specification to yield the exact same results. The key is to communicate when the results hold up. This will consequently help inform the reader what drives the statistical power behind the results.

Broadly, quasi-experimental research aspires to identify effects that do not rely on the underlying assumptions outside of the experimental variation. There are many places where that can break down, including functional form assumptions, external validity, and various confounding effects. The focus is on a robust single causal relationship.

Mechanism: Why Does x Cause y to Change?

The most effective papers typically do not stop with identifying a causal effect and its magnitude. After identifying a likely causal relationship, it is important to assess why x causes y to shift. Understanding mechanisms is often a key goal of social science. There are at least three benefits of establishing mechanisms. First, it provides a rationale for why the effect should exist in the first place. It requires the authors to think about the theoretical contribution of their research more carefully and helps make the argument for causal identification more convincing. Second, identifying mechanisms can help evaluate the benefits and negative consequences of the intervention and identify avenues for course correction, if needed. Third, understanding mechanisms allows for the possibility to extrapolate

the findings to other contexts. Research needs to provide guidance on when and why the causal relationship is relevant.

Assessing the Mechanism Through Mediation Analysis

When the data afford a direct measure of mediator variables, mechanisms can be inferred by mediator analysis. To illustrate how quasi-experiments can show process through mediation, we use Habel, Alavi, and Linsenmayer (2021) as an example. They investigate whether a variable compensation scheme increases salespeople's stress, resulting in emotional exhaustion and more sick days, and counteracts the sales benefits companies might expect from variable compensation schemes. In one of their empirical analyses, they use a natural experiment where a company dropped the variable compensation share from 80% to 20% in one of its business units. To test the health state as a possible mediator variable, they were able to measure sick days both before and after the change in the variable compensation share. In the country of study, sick days are strictly regulated by law and require certification by a physician (at the latest on the third day of the leave). Those who take more than three sick days in a given month are more likely to have substantial health problems. They measure the sick days counting after the third sick day in a month.

Combining the DID analysis with mediator analysis, Habel, Alavi, and Linsenmayer (2021) show that the direct effect of the treatment (drop in variable compensation share) on sales performance is significant and negative, and that the indirect effect of the treatment on sales performance via sick days is positive and significant. The mediator analysis suggests that a higher variable compensation share is associated with enhanced sales performance but also with more sick days, which, in turn, reduce the gains to sales performance.

Assessing the Mechanism Through Moderation Analysis

Heterogeneous treatment effects can be used to test behavioral mechanisms. In a quasi-experimental setting, mechanism checks via heterogeneous treatment effects, sometimes referred to as falsification checks, are not simply equal to identifying moderators. They involve identifying which groups would be affected by a certain mechanism that would display the causal effect of interest, and which other groups would not display the causal effect of interest by the proposed mechanism.

Moderation analysis therefore serves a broader purpose by providing 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. If the effect is larger when theory suggests it should be, then this also helps identify the mechanism. A simple approach is to estimate the effect separately by whether an individual is a member of a group that theory suggests should experience a bigger effect. Formal testing of whether the difference is statistically significant requires a three-way interaction between x , the source of variation, and group membership.

There are many relevant examples in marketing of the use of moderation analyses to demonstrate a mechanism if there is a reason to believe the boundary of underlying process exists or the magnitude of the treatment effect varies by some observables. For example, after showing the European privacy regulation hurt online advertising, Goldfarb and Tucker (2011a) 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. The paper then 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, mechanism checks through mediator or moderation analyses are important because they distinguish the goal of the marketing scholar from the marketing practitioner. Marketing practitioners run experiments and analyze data to understand what they should do in the particular situation they are facing. Marketing scholars need to have a broader sense of applicability beyond the specific setting being studied. Mediation and moderation analyses provide an understanding of when a marketing action will and will not lead to the desired behavior. For this reason, marketing papers are more likely to be remembered for the evidence that is shown in support of a theory explaining why the result holds.

External Validity: How Generalizable is the Effect of x on y ?

The external validity discussion in a paper should recognize the assumptions required for the analysis to capture the ATE across the population of interest, rather than a more local effect that is an artifact of the data sample or the source of quasi-experimental variation. A key concept is the ATE across the entire population. This is the difference in outcomes that would occur by moving the entire population from the control group to the treatment group. However, in some cases, the ATE may not be particularly relevant, because it averages across the entire population and includes units that would never be eligible for treatment (Wooldridge 2000, p. 604). For example, we would not want to include millionaires in computing the ATE of a job training program. To address this, the researcher could use the average treatment effect on the treated, which measures the expected effect of treatment for those who actually were in the treatment condition.

One reason why a research setting may fail to be externally valid is if the treated population is unrepresentative (Lynch 1982). A concern that will drive whether the treated population is unrepresentative is whether those affected could self-select into and out of the treatment. For example, Chiou and Tucker (2012) study a rule change by Google that allowed non-trademark holders to use trademarks in search advertising copy. They study the rule change's effect on user click behavior. In this case, many advertisers did not alter their advertising copy strategy, for a variety of reasons. These advertisers may be

systematically different from the advertisers that did change their strategy. Because these advertisers were not forced to change their strategy, we will never know what would have happened if they did. When faced with such issues, it is best to spell out the potential for self-selection and discuss whether it makes the paper more or less relevant. In this case, it would be accurate to say that the researchers captured the effect of a loosening of trademark restrictions, because it is unlikely that a search engine would force its advertisers into using other advertisers' trademarks. However, it would not be accurate to claim that the researchers capture the broader effect of all advertisers using other advertisers' trademarks in their copy.

The treated population may also be unrepresentative if the treatment impacts a subpopulation to change behavior, but not the main population of interest. This means that the measured effect is localized to that subpopulation, and it is referred to in the literature as the local average treatment effect (LATE). For example, in the context of regression discontinuity, the LATE is the average of the treatment effect over the individuals who would have been in the counterfactual condition if the discontinuity threshold were changed. 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. The idea of LATE also has implications for the interpretation of instrumental variables estimates, as any IV estimate is the LATE for the observations in the regression who experienced the kind of variation exploited by the instrument.⁶

More broadly, as with other aspects of quasi-experimental research, 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 DID 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, Sun and Zhu's article explicitly limits the temptation of the reader to extrapolate too much.

An internally valid quasi-experimental estimate can have broader external validity when used to identify relationships such as elasticities and then to use a structural model to identify the counterfactual of interest. In these cases, under the assumption that the model is a useful representation of reality, quasi-experimental methods serve as a complement for, rather than a substitute to, structure. For example, Anderson et al. (2015) 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, Finkelstein, and Cullen (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, Steenburgh, and Sudhir (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.

Overall, effective quasi-experimental research requires an understanding of the underlying assumptions behind any broad interpretation of quasi-experimental results. 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.

Apologies: What Remains Unproven and What Are the 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.

As an example, Guo, Sriram, and Manchanda (2020) employ a DID research design to study the effect of the payment disclosure law introduced in Massachusetts in June 2009. The research design uses the setting that physicians located in the border counties of Massachusetts and its neighboring states did not have disclosure laws during this period. They lay out the assumptions underlying their estimation:

Our identification of the effect of disclosure legislation relies on the change in new prescriptions by physicians located in Massachusetts (MA) after the policy intervention, relative to their counterparts from "control" states in which no such law existed in the same period.... To assess potential threats to the validity of our research design, we verify if the result was driven by changes in physician payments as a result of the MA disclosure law. If such payment changes were primarily driven by local pharmaceutical reps reallocating their marketing budgets across physicians operating on either side of state borders, this would render the border identification strategy problematic.

(Guo, Sriram, and Manchanda 2020, p. 517)

This example communicates three distinct points. First, it explains the identification strategy. Second, it details the main threats to the validity of this identification strategy. Third, it describes what they do to address it. These points suggest that effective apologies focus on demonstrating what interpretations

⁶ Recent work has shown subtleties in interpreting IV results as LATE (Blandhol et al. 2022).

are reasonable, and what might be a stretch of the results. The goal is not to show that in all circumstances and every conceivable way the identification is perfect. That is not possible. Instead, the goal is to provide clear bounds on the interpretation. The paper's contribution is then a function of whether it provides new knowledge under this bounded interpretation.

Conclusion

Quasi-experimental techniques are an important tool for marketers. First, marketing scholars need to be able to inform marketing practitioners—both managers and policy makers—about the causal effect to allow practitioners to make superior decisions. Second, the best quasi-experimental papers do not simply prove a causal effect but delve into the underlying mechanism, which is key to marketing scholarship's goal of generalizability. Third, such techniques become more important as the scope and span of marketing practice expands and there are new settings and more varied sources of data that allow their application.

The objective of a quasi-experimental research paper is to answer an interesting and important research question about a causal relationship and provide evidence suggesting the mechanism behind the relationship. The choice of method (DID, regression discontinuity, or instrumental variables) depends on the nature of the quasi-experiment. The framework we present focuses on understanding how exogenous variation helps uncover causal relationships and why specific actions affect behavior. Of course the details of the methods will evolve over time as new research appears. Because marketing scholars are often interested in providing generalizable insights about how marketing actions change the behavior of individual consumers, the quasi-experimental framework is particularly useful. Similarly, firms that want to use those insights benefit. As the availability of detailed data grows and marketing technology changes, these methods will enable marketing scholars to provide assessments of a wide variety of situations in which a particular marketing action is likely to change consumer behavior or market dynamics.

Acknowledgments

This article builds on presentations given at the 2013 Workshop on Quantitative Marketing and Structural Econometrics and at the 2012 and 2013 ISMS Doctoral Consortia. The authors thank David Godes, Brett Gordon, Avery Haviv, Joon Ho Lim, Cristina Nistor, and Nathan Yang for comments. They also thank the *JM* review team for their valuable guidance during the review process.

Associate Editors

Christine Moorman and Harald van Heerde


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Catherine Tucker  <https://orcid.org/0000-0002-1847-4832>

References

- Abadie, Alberto (2021), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59 (2), 391–425.
- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge (2017), "When Should You Adjust Standard Errors for Clustering?" *Working Paper 24003*, National Bureau of Economic Research, <https://www.nber.org/papers/w24003>.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105 (490), 493–505.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59 (2), 495–510.
- Akca, Selin and Anita Rao (2020), "Value of Aggregators," *Marketing Science*, 39 (5), 893–922.
- Almond, Douglas, Joseph Doyle, Amanda Kowalski, and Heidi Williams (2010), "Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns," *Quarterly Journal of Economics*, 125 (2), 591–634.
- Altonji, Joseph, Todd Elder, and Christopher Taber (2005), "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113 (1), 151–84.
- Anderson, Eric T., Nathan M. Fong, Duncan I. Simester, and Catherine E. Tucker (2010), "How Sales Taxes Affect Customer and Firm Behavior: The Role of Search on the Internet," *Journal of Marketing Research*, 47 (2), 229–39.
- Anderson, Soren T., Ryan Kellogg, Ashley Langer, and James M. Sallee (2015), "The Intergenerational Transmission of Automobile Brand Preferences," *Journal of Industrial Economics*, 63 (4), 763–93.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91 (434), 444–55.
- Angrist, Joshua D. and Michal Kolesár (2021), "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," *Working Paper 29417*, National Bureau of Economic Research, <https://www.nber.org/papers/w29417>.
- Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Arkhangelsky, Dmitry, Susan Athey, David Hirshberg, Guido W. Imbens, and Stefan Wager (2021), "Synthetic Difference-in-Differences," *American Economic Review*, 111 (12), 4088–4118.

- Athey, Susan and Guido W. Imbens (2022), "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 226 (1), 62–79.
- Bareca, Alan I., Melanie Guldi, Jason M. Lindo, and Glen R. Waddell (2011), "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *Quarterly Journal of Economics*, 126 (4), 2117–23.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119 (1), 249–75.
- Blake, Thomas, Chris Nosko, and Steven Tadelis (2015), "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 83 (1), 155–74.
- Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2022), "When Is TSLS Actually LATE?" Technical Report 29709, National Bureau of Economic Research.
- Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022), "Quasi-Experimental Shift-Share Research Designs," *Review of Economics Studies*, 89 (1), 181–213.
- Bronnenberg, Bart, Jean-Pierre Dubé, and Matthew Gentzkow (2012), "The Evolution of Brand Preferences: Evidence from Consumer Migration," *American Economic Review*, 102 (6), 2472–2508.
- Callaway, Brantly and Pedro H.C. Sant'Anna (2020), "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225 (2), 200–30.
- Cameron, Colin, Jonah Gelback, and Douglas Miller (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90 (3), 414–27.
- Campbell, Roy H. (1965), "A Managerial Approach to Advertising Measurement," *Journal of Marketing*, 29 (4), 1–6.
- Canay, Ivan A., Andres Santos, and Azeem M. Shaikh (2021), "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *Review of Economics and Statistics*, 103 (2), 346–63.
- Cattaneo, Matias D., Rocio Titiunik, and Gonzalo Vazquez-Bare (2020), "The Regression Discontinuity Design," in *Handbook of Research Methods in Political Science and International Relations*, Luigi Curini and Robert Franzese, eds. SAGE Publications, 835–57.
- Chandy, Rajesh K., Gita Venkataramani Johar, Christine Moorman, and John H. Roberts (2021), "Better Marketing for a Better World," *Journal of Marketing*, 85 (3), 1–9.
- Chevalier, Judith and Dina Mayzlin (2006), "The Effect of Word of Mouth Online: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345–54.
- Chiou, Lesley and Catherine E. Tucker (2012), "How Does the Use of Trademarks by Third-Party Sellers Affect Online Search?" *Marketing Science*, 31 (5), 819–37.
- Chung, Doug J., Thomas Steenburgh, and K. Sudhir (2014), "Do Bonuses Enhance Sales Productivity? A Dynamic Structural Analysis of Bonus-Based Compensation Plans," *Marketing Science*, 33 (2), 165–87.
- Conley, Timothy G. and Christopher R. Taber (2011), "Inference with 'Difference in Differences' with a Small Number of Policy Changes," *Review of Economics and Statistics*, 93 (1), 113–25.
- Cook, Thomas D. and Donald T. Campbell (1979), *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.
- Crayton, Ancil (2020), "Causal Inference for Data Scientists: Econometrics to Machine Learning," (accessed January 26, 2022), <https://www.ancilcrayton.com/talk/ct-2020/>.
- Datta, Hannes, George Knox, and Bart Bronnenberg (2018), "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery," *Marketing Science*, 37 (1), 5–21.
- Deaton, Angus S. (2009), "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," *Proceedings of the British Academy*, 162, 123–60.
- De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110 (9), 2964–96.
- Deighton, John A., Carl F. Mela, and Christine Moorman (2021), "Marketing Thinking and Doing," *Journal of Marketing*, 85 (1), 1–6.
- DiPrete, Thomas A. and Markus Gangl (2004), "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments," *Sociological Methodology*, 34 (1), 271–310.
- Donald, Stephen and Kevin Lang (2007), "Inference with Difference in Differences and Other Panel Data," *Review of Economics and Statistics*, 89 (2), 221–33.
- Doudchenko, Nikolay and Guido W. Imbens (2016), "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research, <https://www.nber.org/papers/w22791>.
- Dubé, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi (2018), "Income and Wealth Effects on Private-Label Demand: Evidence from the Great Recession," *Marketing Science*, 37 (1), 22–53.
- Ebbes, Peter, Dominik Papies, and Harald J. Van Heerde (2011), "The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity," *Marketing Science*, 30 (6), 1115–22.
- Einav, Liran, Amy Finkelstein, and Mark Cullen (2010), "Estimating Welfare in Insurance Markets Using Variation in Prices," *Quarterly Journal of Economics*, 125 (3), 877–921.
- Elberg, Andres, Pedro M. Gardete, Rosario Macera, and Carlos Noton (2019), "Dynamic Effects of Price Promotions: Field Evidence, Consumer Search, and Supply-Side Implications," *Quantitative Marketing and Economics*, 17 (1), 1–58.
- Gelman, Andrew (2010), "Experimental Reasoning in Social Science," in *Field Experiments and Their Critics*, Chap. 7. New Haven, CT: Yale University Press, 185–95.
- Gibbons, Charles E., Juan Carlos Suarez Serrato, and Michael B. Urbancic (2017), "Broken or Fixed Effects?" *Journal of Econometric Methods*, 8 (1), 20170002.
- Gill, Manpreet, Shrihari Sridhar, and Rajdeep Grewal (2017), "Return on Engagement Initiatives: A Study of a Business-to-Business Mobile App," *Journal of Marketing*, 81 (4), 45–66.
- Goldfarb, Avi and Catherine E. Tucker (2011a), "Privacy Regulation and Online Advertising," *Management Science*, 57 (1), 57–71.
- Goldfarb, Avi and Catherine E. Tucker (2011b), "Search Engine Advertising: Channel Substitution When Pricing Ads to Context," *Management Science*, 57 (3), 458–70.
- Goodman-Bacon, Andrew (2021), "Difference-in-Difference with Variations in Treatment Timing," *Journal of Econometrics*, 225 (2), 247–77.
- Grewal, Rajdeep (2017), "Journal of Marketing Research: Looking Forward," *Journal of Marketing Research*, 54 (1), 1–4.

- Guo, Tong, Srinivasaraghavan Sriram, and Puneet Manchanda (2020), "Let the Sunshine In: The Impact of Industry Payment Disclosure on Physician Prescription Behavior," *Marketing Science*, 39 (3), 516–39.
- Habel, Johannes, Sascha Alavi, and Kim Linsenmayer (2021), "Variable Compensation and Salesperson Health," *Journal of Marketing*, 85 (3), 130–49.
- Hagemann, Andreas (2019), "Placebo Inference on Treatment Effects when the Number of Clusters Is Small," *Journal of Econometrics*, 213 (1), 190–209.
- Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69 (1), 201–9.
- Hartmann, Wesley and Daniel Klapper (2018), "Super Bowl Ads," *Marketing Science*, 37 (1), 78–96.
- Hartmann, Wesley, Harikesh S. Nair, and Sridhar Narayanan (2011), "Identifying Causal Marketing Mix Effects Using a Regression Discontinuity Design," *Marketing Science*, 30 (6), 1079–97.
- Heckman, James J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46 (7/4), 931–59.
- Heckman, James J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, 115 (1), 47–97.
- Hollenbeck, Brett, Sridhar Moorthy, and Davide Proserpio (2019), "Advertising Strategy in the Presence of Reviews: An Empirical Analysis," *Marketing Science*, 38 (5), 793–811.
- Imbens, Guido W. and Thomas Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–35.
- Imbens, Guido W. and Donald B. Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge University Press.
- Imbens, Guido W. and Jeffrey Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47 (1), 5–86.
- Israeli, Ayelet (2018), "Online MAP Enforcement: Evidence from a Quasi-Experiment," *Marketing Science*, 37 (5), 710–32.
- Janakiraman, Ramkumar, Joon Ho Lim, and Rishika Rishika (2018), "The Effect of a Data Breach Announcement on Customer Behavior: Evidence from a Multichannel Retailer," *Journal of Marketing*, 82 (2), 85–105.
- Kolesár, Michal and Christoph Rothe (2018), "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108 (8), 2277–2304.
- Lambrecht, Anja, Katja Seim, and Catherine E. Tucker (2011), "Stuck in the Adoption Funnel: The Effect of Interruptions in the Adoption Process on Usage," *Marketing Science*, 30 (2), 355–67.
- Lancaster, Tony (2000), "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95 (2), 391–413.
- Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack R. Porter (2021), *Valid T-Ratio Inference for IV*. Cambridge, MA: National Bureau of Economic Research.
- Lim, Joon Ho, Rishika Rishika, Ramkumar Janakiraman, and P.K. Kannan (2020), "Competitive Effects of Front-of-Package Nutrition Labeling Adoption on Nutritional Quality: Evidence from Facts Up Front–Style Labels," *Journal of Marketing*, 84 (6), 3–21.
- List, John A. (2011), "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off," *Journal of Economic Perspectives*, 25 (3), 3–16.
- Luca, Michael (2011), "Reviews, Reputation, and Revenue: The Case of Yelp.com," Harvard Business School Working Papers 12-016, Harvard Business School.
- Lynch, John G. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (3), 225–39.
- Manchanda, Puneet, Grant Packard, and Adithya Pattabhiramaiah (2015), "Social Dollars: The Economic Impact of Customer Participation in a Firm-Sponsored Online Customer Community," *Marketing Science*, 34 (3), 367–87.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421–55.
- Meyer, Bruce (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics*, 12 (2), 151–62.
- Mooman, Christine, Rosellina Ferraro, and Joel Huber (2012), "Unintended Nutrition Consequences: Firm Responses to the Nutrition Labeling and Education Act," *Marketing Science*, 31 (5), 717–37.
- Oster, Emily (2019), "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business and Economic Statistics*, 37 (2), 187–204.
- Pattabhiramaiah, Adithya, S. Sriram, and Puneet Manchanda (2019), "Paywalls: Monetizing Online Content," *Journal of Marketing*, 83 (2), 19–36.
- Qian, Yi (2008), "Impacts of Entry by Counterfeiters," *Quarterly Journal of Economics*, 123 (4), 1577–1609.
- Ramani, Nandini and Raji Srinivasan (2019), "Effects of Liberalization on Incumbent Firms' Marketing-Mix Responses and Performance: Evidence from a Quasi-Experiment," *Journal of Marketing*, 83 (5), 97–114.
- Rebecq, Antoine (2020), "Causal Inference Cheat Sheet for Data Scientists," Towards Data Science (April 29), <https://towardsdatascience.com/causal-inference-cheat-sheet-for-data-scientists-a1d97b98d515>.
- Rosenbaum, Paul R. (2002), *Observational Studies*, 2nd ed. New York: Springer.
- Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 71 (1), 41–55.
- Rossi, Peter E. (2014), "Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications," *Marketing Science*, 33 (5), 621–762.
- Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe (2022), "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," arXiv preprint arXiv:2201.01194.
- Rubin, Donald B. (2005), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100 (469), 322–31.
- Schmitt, Bernd H., June Cotte, Markus Giesler, Andrew T. Stephen, and Stacy Wood (2021), "Our Journal, Our Intellectual Home," *Journal of Consumer Research*, 47 (5), 633–35.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin Company.

- Shapiro, Bradley T. (2020), "Advertising in Health Insurance Markets," *Marketing Science*, 39 (3), 587–611.
- Shin, Jiwoong, K. Sudhir, and Dae-Hee Yoon (2012), "When to Fire Customers: Customer Cost-Based Pricing," *Management Science*, 58 (5), 932–47.
- Shriver, Scott K., Harikesh Nair, and Reto Hofstetter (2013), "Social Ties and User Generated Content: Evidence from an Online Social Network," *Management Science*, 59 (6), 1425–43.
- Stock, James, Jonathan Wright, and Motohiro Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20 (4), 518–29.
- Sun, Liyang and Sarah Abraham (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225 (2), 175–99.
- Sun, Monic and Feng Zhu (2013), "Ad Revenue and Content Commercialization: Evidence from Blogs," *Management Science*, 59 (10), 2314–31.
- Thomas, Michael (2020), "Spillovers from Mass Advertising: An Identification Strategy," *Marketing Science*, 39 (4), 807–26.
- Toubia, Olivier (2022), "Editorial: A New Chapter or a New Page for Marketing Science?" *Marketing Science*, 41 (1), 1–6.
- Tucker, Catherine E., Juanjuan Zhang, and Ting Zhu (2013), "Days on Market and Home Sales," *RAND Journal of Economics*, 44 (2), 337–60.
- Van Heerde, Harald J., Christine Moorman, C. Page Moreau, and Robert W. Palmatier (2021), "Reality Check: Infusing Ecological Value into Academic Marketing Research," *Journal of Marketing*, 85 (2), 1–13.
- Wang, Yanwen, Michael Lewis, and David Schweidel (2018), "A Border Strategy Analysis of Ad Source and Message Tone in Senatorial Campaigns," *Marketing Science*, 37 (3), 333–55.
- Wang, Yanwen, Chunhua Wu, and Ting Zhu (2019), "Mobile Hailing Technology Value and Taxi Driving Behaviors," *Marketing Science*, 38 (5), 733–912.
- Wooldridge, Jeffrey (2000), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- Zhang, Juanjuan (2010), "The Sound of Silence: Observational Learning in the U.S. Kidney Market," *Marketing Science*, 29 (2), 315–35.