Unpacking the Instrumental Variables Approach*

Rajdeep Grewal
Yeşim Orhun

1 Introduction

Marketing scholars and practitioners are frequently encounter causal questions related to strategic marketing decisions. Examples of such decisions include pricing, advertising, market entry, product development, brand positioning, contractual choices, and distribution decisions, to name a few. The factors and rules shaping strategic marketing decisions are often not fully observed by the researcher. When unobserved factors are also associated with the outcome of interest, this confounding relationship, known as the omitted variables or common causes issue, hinders the identification of the causal relationship of interest. For example, when estimating a causal impact of advertising spending on demand, researchers may find that both advertising spend and sales of a product are driven by the product's inherent market potential. Unobserved confounders create an identification challenges that empirical marketing scholars and practitioners often face when answering causal questions related to marketing decisions.

While randomized controlled experiments may be viable in some cases to address this issue, for many important strategic decisions, experimentation may not be feasible or ethical. In such cases, researchers use quasi-experimental approaches. Goldfarb et al. (2022) provide an excellent overview and detailed guidance for a variety of quasi-experimental methods, including the instrumental variable (IV) approach. Papies et al. (2023) review the history of marketing literature on the use of various quasi-experimental methods and identify instrumental variables as the most common method.

Despite the broad use of the IV approach, the nature of confoundedness in a particular empirical context and how the proposed IV addresses it is, unfortunately, not always as clear as it could be. In this piece, we build on existing work to provide marketing scholars and practitioners with a resource that we hope will help them (1) identify confoundedness

*The authors thank Elea Feit, Avi Goldfarb, and Xiao Liu for their helpful feedback.

concerns in their empirical context, (2) transparently discuss the assumptions that need to hold for the IV approach to be valid in uncovering the causal effect, and (3) evaluate the plausibility of these assumptions.

We hope researchers use this piece not as a checklist but to *critically* assess whether the IV approach is appropriate given their research question and data. The IV approach has been used across markedly different empirical contexts and to answer a wide range of causal questions in the marketing literature. For example, recent articles using the IV approach examine the impact of review variance on demand (Lee et al., 2023), the role of television advertising in satellite operators' commercial success (Yang et al., 2021), the influence of pictures on the engagement with social media posts (Li and Xie, 2020), and whether money-back guarantees can serve as a signal of quality (Yu et al., 2022), to name a few. Each of these questions and empirical contexts presents a unique identification challenge, and the validity of the IV approach in each case has to be established based on a unique set of facts and arguments. Papies et al. (2023) note, "One of the main lessons from history is that there are no easy turn-key solutions to an endogeneity problem. Each of the methods used to address endogeneity in observational data relies on assumptions. It is critical that researchers using these methods carefully assess these assumptions in the context of their research question and data." We hope that this piece will aid the readers in this regard.

2 The Setup

Figure 1 presents a directed acyclic graph (DAG) demonstrating the omitted variable bias problem arising from unobserved confounders.[1] The causal impact we want to identify is the influence of treatment D on outcome Y ($D \to Y$). For example, we might be interested in the impact of price (D) on sales (Y) for diet soda. The observed common causes W and the unobserved common causes U (unobservability to the researchers is represented by dashed lines) impact both D and Y. These sets of variables are referred to as *confounders* for the direct causal relationship of interest, because they lead to an association between D and Y even in the absence of a causal relationship. For example, we might imagine that a manager's sales expectations may be associated with both the pricing decision and the sales outcomes. Typically, expectations are unknown (an example of U). The price also responds to input costs, which may also be associated with demand (an example of W if observed). For example, in the case of diet soda, aspartame prices may go up when the demand for the

[1] In a directed graph, each node is a random variable, and the edges are directed, indicating causal associations in the direction of the arrow. In acyclic graphs, causality runs in one direction. To learn more about DAGs and their usefulness, see Pearl (2009) and Imbens (2020).

diet soda category increases. It is relatively straightforward to deal with confounding due to W by conditioning on W, since these variables are observable.[2] The main identification challenge to uncover the causal impact $D \rightarrow Y$ arises from the existence of unobserved confounders U.

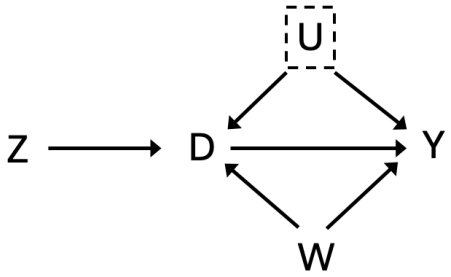

Figure 1: Directed Acyclic Graph representation of the omitted variable bias and IV approach.

Figure 1 also includes a variable Z, which has a mediated pathway from Z to Y. In the case of a randomized controlled trial, Z indicates the random experimental assignment to treatment and control arms. Alternatively, Z can refer to an IV, which is not experimentally randomized, but is "as good as random" for the purposes of identifying $D \rightarrow Y$. Either way, if five assumptions, which we detail below, are satisfied on the causal chain from Z to Y, Z can be used to identify the $D \rightarrow Y$ relationship.

3 Assumptions

For Z to be a valid IV for identifying the causal effect of $D \rightarrow Y$, five assumptions need to be satisfied: (1) independence assumption, (2) stable unit treatment value assumption (SUTVA), (3) inclusion restriction, (4) exclusion restriction, and (5) monotonicity assumption. The plausibility of these assumptions determine whether the IV approach is suitable for a specific research question within a given institutional context. Below, we explain these assumptions and detail their implications in the context of an example to build intuition. For a more technical coverage, we recommend readers consult econometrics textbooks and articles, such as Angrist et al. (1996), Angrist and Krueger (2001), Angrist and Pischke (2009), Cunningham (2021), and Murray (2006), among others.

[2]Care must be taken when controlling for the $W \rightarrow Y$ relationship. A model with strong functional form assumptions, such as a linear regression model, may not be appropriate. Approaches such as matching, nonparametric methods, and machine learning methods can be used instead.

3.1 Five Assumptions

The first assumption, called the **independence assumption**, also known as the ignorability or exchangeability assumption, stipulates that the assignment of different levels of Z to different units (e.g., firms, individuals) is as good as random. In our DAG, this means that Z is not associated with U or W (indicated by the fact that no edge is drawn between Z and these variables in Figure 1). We can weaken this assumption to conditional independence. For example, if there were an association between Z and W, conditional independence could be achieved by controlling for W.[3] One can alternatively think of this assumption as stipulating that either there is no selection into Z, and if there is selection, it is only on observables. The researcher needs to defend the plausibility of this assumption, which is most easily done when the instrument Z is indeed randomized. Otherwise, why believe the conditional independence of Z, but not D? The answer will depend on the institutional context, as we discuss below.

The second assumption, **stable unit treatment value assumption (SUTVA)**, states that the value of unit i's instrument or treatment does not affect other units' potential outcomes, i.e., there are no unmodeled spillovers or interference. This assumption permits us to write $D_i(Z) = D_i(Z_i)$ and $Y_i(Z, D(Z)) = Y_i(Z_i, D_i(Z_i))$. This is not a trivial assumption and researchers should contemplate its plausibility carefully. In practice, SUTVA violations can occur due to many reasons, including general equilibrium effects, anticipation, contagion, information spillovers, social comparisons, externalities, and network effects, among others. Even in the context of a randomized experiment, SUTVA may be violated. For example, if an instrument (or experimental manipulation) varies the prices of a subset of firms, it can also have an impact on the prices of untreated firms through competitive pricing (e.g., see Holtz et al., 2024).

When the SUTVA and independence assumptions are satisfied, the IV is referred to as being *exogenous*. Many marketing papers focus on the idenpendence assumption and fail to consider SUTVA in discussing the validity of their IV. Moreoever, it is common for marketing papers to refer to the independence assumption as the exclusion restriction - a confusion that is possibly driven by the necessity of the independence assumption in establishing exogeneity. We discuss the exclusion restriction below and highlight the need to discuss the plausibility of all five assumptions if the objective is to identify the causal effect $D \to Y$.

It is important to note that the independence and SUTVA assumptions are sufficient for $Z \to Y$ to be interpreted causally. In the IV literature, this causal effect is called the *reduced*

[3]With all other assumptions satisfied, the 2SLS with covariates estimates a weighted average of the covariate-specific local average treatment effects (LATEs). Abadie (2003) shows how to estimate the overall LATE using a weighting approach based on a "propensity score" for the instrument.

form effect. Sometimes, the reduced form effect suffices when the research question can be satisfactorily answered by identifying the causal effect of Z on Y.[4] However, if the causal effect of interest is $D \rightarrow Y$, three additional assumptions we discuss below are necessary for the identification of the local average treatment effect (LATE), which is defined as the causal effect $D \rightarrow Y$ among units for which Z had an effect on D. The the IV approach can only identify the "local" effect of D on Y for units whose treatment status can be manipulated by the instrument, and not the average treatment effect among all units. This has implications for generalizability of the identified causal effect, which we return to in our concluding remarks.

The third assumption, called the **inclusion restriction**, also referred to as the first-stage assumption or relevance assumption, is that the instrument Z must influence D. This assumption is the only one among the five assumptions that can be empirically verified. It is expected that papers provide empirical evidence for the association between Z and D, and assess the strength of that relationship by testing the coefficient on the instrument in the so-called first-stage regression of the treatment on the instrument. We return to the issue of power of the first-stage and inference with weak instruments in our concluding remarks.

The fourth assumption, the **exclusion restriction** assumption, indicates that any effect of the instrument Z on outcome Y is exclusively through its effect on D. This assumption is depicted by Figure 1 indicating that there is no other impact of Z on Y once we condition on the value of D. If the model is over-identified, i.e., there are more instruments than endogenous variables, then an overidentification test (e.g., Sargan-Hansen test) can be performed to test whether all instruments are uncorrelated with the 2SLS residuals. However, these tests require a constant effects assumption that is often difficult to defend, and therefore, overidentification tests are not commonly used (e.g., see Kiviet and Kripfganz, 2021). Instead, researchers establish logical support for this assumption using institutional details and falsification tests, as we discuss below.

As we noted earlier, marketing studies sometimes confuse the exclusion restriction with the independence assumption. Thus, it is important to highlight the differences. While the independence assumption is about the instrument (Z) not being correlated with unobserved confounders (U), the exclusion restriction states that the instrument does not impact the dependent variable (Y) except through its effect on the endogenous independent variable (D). For example, in using a cost shifter as an IV for price in estimating the impact of prices

[4]The IV framework parallels the failure-to-treat issue in randomized experiments. We can think of Z as the randomized experimental manipulation, and D as the indicator of whether a person is treated if assigned to treatment. In this interpretation, the reduced form effect ($Z \rightarrow Y$) is the intent-to-treat (ITT) effect (see, for example, Gerber and Green (2012)). Marketing researchers using experiments often only report the impact of the manipulation on the outcome variable, which is the reduced form effect.

on demand, the cost shifter satisfies the exclusion restriction if the only effect that the cost shifter has on demand operates through its effect on price. Thus, not all cost shifters will satisfy the exclusion restriction, a point we will return to when discussing the role of the institutional context in assessing the plausibility of the assumptions.

The fifth assumption, called the **monotonicity assumption**, is the final assumption necessary in the identification of the LATE when $D \rightarrow Y$ is heterogeneous across units (Imbens and Angrist, 1994; Heckman and Vytlacil, 1999; 2000; 2005; Heckman et al., 2006).[5] The monotonicity assumption requires that a hypothetical change in Z either has no impact on a unit's treatment status D, or changes it in the same direction as it does for all other units on which it has an impact. Let's consider a binary Z and a binary D for illustrative purposes. In the language of the potential outcomes model (Rubin, 2005), *compliers* are units whose behavior is impacted by the instrument. For compliers, assume $D = 1$ when $Z = 1$ and $D = 0$ when $Z = 0$. *Defiers* are those for whom the instrument has the opposite effect as compliers: for them, then, $D = 1$ when $Z = 0$ and $D = 0$ when $Z = 1$. *Always-takers* and *never-takers* are not impacted by the instrument: for always-takers, $D = 1$, and for never takers, $D = 0$, regardless of Z. Thus, always-takers and never-takers do not inform the IV estimate. The monotonicity assumption indicates that we can have either compliers or defiers, but not both.[6] The monotonicity assumption would be violated, for example, when a nudge (e.g., anti-smoking ads) works in the expected direction for some but causes a backlash reaction for others. Without making further assumptions, it is not possible to empirically verify the monotonicity assumption.[7]

3.2 A Demonstration

Authors need to translate the implications of these five assumptions into their empirical context and convince the reader of their plausibility. To demonstrate, we discuss the example offered by Calder-Wang and Gompers (2021), who study the impact of employee gender diversity on venture capital firm performance. To identify this causal effect, the paper uses the sex of venture capital partners' children as an instrument for the decision to hire a woman. The **inclusion restriction** conjecture is that partners who have more daughters

[5] If the homogeneity of $D \rightarrow Y$ assumption can be upheld, then (1) we do not need the monotonicity assumption, and (2) the causal effect can be generalized to the population at large, giving us the average treatment effect (ATE), instead of LATE. However, the homogeneity assumption is unlikely to hold in many marketing applications; therefore, we discuss the monotonicity assumption in the main text.

[6] This nomenclature is the reason why the LATE is also referred to Complier Average Causal Effect (CACE). See De Chaisemartin (2017) for inference in the IV approach without the monotonicity assumption.

[7] Frandsen et al. (2023) offer a test of the monotonicity assumption in a particular IV design by making additional assumptions on the average treatment effect among those who violate the monotonicity assumption.

are more likely to hire women. The authors discuss conceptual reasons for this relationship and also present empirical evidence of the first-stage, which we recommend all papers using the IV approach do.

In this context, **SUTVA** requires that the hiring decisions and financial performance of a partner are impacted only by the number of daughters the partner has, and not by the gender of the children of other partners. This assumption can be violated in several ways. For example, if the supply of qualified female employees is extremely tight, then the increased interest in hiring female employees due to the gender composition of one firm's partners could impact the hiring and/or financial performance of a competing firm.

The **independence** assumption is equivalent to assuming that whether partners have sons or daughters (conditional on having children) is as good as random. If certain parents (e.g., those who hold different gender views) employed a gender-based stopping rule (e.g., keep having children until they have at least one son), the independence assumption would be violated. To defend against this particular concern, the authors provide evidence that a first-born daughter does not predict the total number of children, which constitutes an example of a falsification test – a concept we discuss below.

The **exclusion restriction** in this example necessitates that the genders of partners' children do not have an impact on the venture capital firm's performance *except through* the impact on gender diversity in the firm. The authors recognize that if having more daughters directly improves a partner's skills in a way that increases their ability to source or close deals, this assumption would be violated. They provide evidence that venture capital partners with more daughters do not have more successful deals, which is another example of a falsification test.

Finally, the **monotonicity** assumption necessitates that we can only have partners for whom having more daughters would either increase their likelihood of hiring women or not impact it, but we cannot have partners for whom having more daughters would decrease the likelihood of hiring women. This assumption would be violated, for example, if for a minority of partners, having daughters reinforce sexist views of the workplace.

We summarize this discussion in Table 1. The table also includes Sinkinson and Starc (2019) as another working example to demonstrate the assumptions required in using political advertising cycles as an instrument for advertising spend. For a more detailed discussion of potential violations of the exclusion restriction and monotonicity assumption in using political cycles as instruments, and the important role time and market fixed effects play, see Moshary et al. (2021). As these examples make clear, the plausibility of the identifying assumptions needs to be defended with institutional details and supporting empirical patterns. If the assumptions cannot be credibly defended, researchers should not use the IV approach.

4 Evaluating and Defending the Plausibility of Assumptions

None of the identifying assumptions we discussed above, except for the inclusion restriction, can be empirically validated. Therefore, they must be logically established and defended based on common sense, subject matter arguments, and institutional details. Goldfarb et al. (2022) suggest that "... 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 [...]. The credibility of any quasi-experimental work therefore relies on the plausibility of the argument for causality rather than on any formal statistical test." In our assessment, many of the published papers in marketing using the IV approach do not offer a detailed enough discussion of the implications and plausibility of the required assumptions in the empirical context they study. When they do, the discussion tends to focus mainly on the relevance and independence assumptions. In addition to providing a discussion of other assumptions, we suggest researchers treat identification as a central part of the the manuscript's narrative, using institutional details and theory to tie together elements that make the research question important and the identification valid. Thus, in this section, we discuss approaches researchers can take to evaluate the plausibility of all the assumptions in the IV framework to develop a cohesive and convincing story.

Institutional details Subject matter arguments based on institutional knowledge are paramount to judging whether the required assumptions are plausible, and consequently, whether causal inference can be achieved. In many quasi-experimental papers, identifying assumptions are justified solely by subject-matter arguments that use institutional details. The clarity with which the authors translate the identifying assumptions to their context and provide detailed discussion of institutional details that help the reader evaluate these assumptions go a long way in convincing the reader of a causal relationship.

Institutional context makes or breaks an instrument. An instrument that satisfies an assumption naturally in some contexts, may blatantly violate it in others. To illustrate, let's consider cost-based instruments, which are commonly used in marketing and industrial organization to obtain the causal impact of prices on demand. Do input costs as an instrument satisfy the exclusion restriction? The answer depends on the institutional details. For example, consider using orange wholesale prices as an instrument for orange juice prices when estimating demand for orange juice in Michigan. Imagine that a drought in Florida pushed up orange wholesale prices across the nation. It might be relatively straightforward to defend that this input cost variation is plausibly exogenous to demand conditions for

orange juice in Michigan. Now, instead, consider the cost of steel as an input for automobile manufacturing. The automotive industry accounts for 10-15 percent of global steel use and automobile production levels are known to impact steel prices. Therefore, the demand for automobiles may have an impact on steel prices. Alternatively, both steel prices and demand for automobiles may be impacted by the strength of the economy. In this context, it may be hard to refute that steel prices have no association with consumer demand for automobiles except through their impact on car prices.

Another illustration of the institutional context mattering for the validity of an instrument comes from examiner designs (also called judge fixed effect design, or leniency design). In these designs, there is an examiner who has discretion in determining the outcomes (e.g., a judge in a hearing, a grader in a class, or a consumer in responding to a satisfaction survey), and there is systematic heterogeneity in their judgments (e.g., some judges being systematically more lenient than others, some consumers being generally more grumpy). Similar designs have been adapted to study marketing questions (e.g., Li and Xie, 2020; Lee et al., 2023). In cases where the assignment of the examiner is as good as random, we can consider the identity of the examiner as an instrument for the treatment whose effect we are trying to examine. For example, consider being interested in the impact of review valence on product sales, and assume that certain consumers are systematically more negative in their review behavior and that the arrival of consumer types (in terms of their overall negativity) is random. We could imagine using the identity of the consumer as an instrument for review valence. In this context, is the monotonicity assumption satisfied? It depends. Monotonicity holds whenever any product that would have received a 4-star rating from a generally negative consumer receives a 4-star or 5-star rating from a happy-go-lucky consumer. It is violated if the happy-go-lucky consumer sometimes rates products worse than the generally negative consumer. In this context, the researcher may argue that the assumption is more likely to hold within a product category, rather than across categories. However, even within a product category, consumers may vary in what makes them unhappy. For example, the happy-go-lucky consumer might be generally positive, except in cases where the product arrives damaged, in which case they switch and become even more negative than the generally negative consumer. The institutional context matters greatly in the researcher's ability to make a case for (or refute) the likelihood of this scenario.[8]

There are many papers that bring institutional details expertly into the narrative and use them to lay out the rationale for the plausiblity of the identifying assumptions. For example,

[8]For more details on inference in examiner designs, we refer the interested reader to Chyn et al. (2024). Chapter 7.8.2 of Cunningham (2021) also provides a useful discussion of the plausibility of the independence, exclusion restriction, and monotonicity assumptions in examiner designs.

consider Bruhn et al. (2022), who examine whether veterans who have experienced more combat exposure are more likely to have negative life outcomes post-deployment (education, financial heath, suicide, incarceration, etc.). They clearly explain the institutional process of brigade assignments in the Army to support the relevance of their instrument and defend the (conditional) independence assumption. In making a case for the plausibility of the exclusion restriction against one potential criticism, Sinkinson and Starc (2019) point out that detailing spending levels are set at the annual level and therefore cannot be quickly adapted at the market level in response to TV ad spending declining. Similarly, in examining peer effects on salesperson quitting behavior using the IV approach Sunder et al. (2017) argues that the management's evaluation of the salesperson in the first month of them joining the firm satisfies the independence assumption due to its private nature. We recommend that all authors think through the institutional details when specifying a causal model and picking an instrument, and communicate these details to their readership in the context of the five identifying assumptions we discussed above.

Falsification Tests An important benefit of specifying the identifying assumptions of a causal research designs is that these assumptions often have falsifiable implications. Although identifying assumptions cannot be verified empirically, researchers can conduct tests for these implications to check whether they can be empirically refuted (Angrist and Krueger, 1999). These tests are called falsification tests. Failed falsification attempts do not prove the assumption they are designed to falsify, but can help build the case for the plausibility of the identifying assumptions. In this section, we aim to build intuition for coming up with falsification tests by sharing examples of tests used by researchers across a variety of areas.

In devising a falsification test for the independence assumption, researchers often check for balance in observables across levels of Z. The idea here is that if Z is as-good-as-random, then we would not expect any systematic differences in the means or distributions, of pre-treatment covariates across different levels of Z.[9] At other times, the researchers evaluate a plausible threat to the independence assumption. We discussed one such example in the case of Calder-Wang and Gompers (2021) above. Another example refuting a threat to the independence assumption comes from Gong et al. (2017). The authors conduct a field experiment to show that tweets by a media company on Sina Weibo about their TV shows increase the viewership of these shows. The experiment assigns a subset of shows to receive company tweets. The authors want to refute the possibility that the assignment of shows was somehow correlated with unobservables that drive show viewership (e.g., show popularity).

[9]Of course, if the assumption is conditional independence, the balance assessment is also conditional. When assessing balance across a number of variables, instead of running numerous independent balance tests, it is better to employ one omnibus balance test (e.g., Hansen and Bowers, 2008).

To do so, they exploit the fact that tweeting happens in a single point in time, but the TV show airs at different times across geographies. Under the independence assumption, we would not expect any impact of treatment when tweeting happened after the show airs in a geography. If the assignment correlated with unobservables, however, we would expect a higher viewership of shows in geographies even if tweeting happened after the show airs. This constitutes a clever falsification test of the independence assumption in the paper's institutional context.

To create a falsification test for the exclusion restriction, researchers ask what empirical pattern might suggest that Z has an impact on Y that did not operate through D alone. The falsification test Calder-Wang and Gompers (2021) offered for the exclusion restriction was an example of evaluating a particular channel by which Z may have a direct effect on Y. In most cases, falsifying the exclusion restriction involves testing the reduced form effect of Z on Y in situations where it is impossible (or, extremely unlikely) for Z to influence D. If Z has an impact on Y in these situations, it would show that Z's impact on Y does not always operate through D. Sometimes, it may be possible to evaluate the reduced form effect among never-takers as a falsification test. For example, Erikson and Stoker (2011) use the Vietnam draft lottery based on birth dates as an instrument for vulnerability to military service to study whether this vulnerability impacts political attitudes. Under the exclusion restriction assumption, birth dates should not affect (1) men's attitudes who are exempt from the draft due to college deferrals, or (2) women's attitudes. Thus, the reduced form effect is expected to be zero for these two groups of never-takers, unless birth dates have a direct effect on political attitudes that does not only operate through vulnerability to military service.

In other cases, it might be possible to devise falsification tests based on the fact that the exclusion restriction would predict a null effect of the instrument among the always-takers. For example, in the context of medical research, a physician's general tendency to operate is used as an instrument for whether patients received surgery, in order to examine the impact of having surgery versus not having it on mortality. Yang et al. (2014) provide a falsification test that exploits the fact that certain subgroups of patients in the data are always operated on, and therefore their in-hospital mortality should not be affected by their physician's general tendency to operate (i.e., the reduced form effect among always-takers should be zero) if the exclusion restriction holds (also, see Keele et al., 2019).

Falsification tests for other assumptions can also be devised based on the particulars of the empirical context (see, e.g., Burke et al., 2019; Danieli et al., 2023). Overall, falsification tests can be a useful tool in making the case for the plausibility of the identifying assumptions, but as we cautioned above, should not be interpreted as providing proof for them.

5 Concluding Remarks

In this piece, we provided a brief discussion of the identifying assumptions in the IV approach, focusing on the importance of assessing their plausibility in a given empirical context, and possessing sufficient institutional knowledge to do so. While we focused on one approach, we hope that many of the insights are useful for researchers in thinking through other quasi-experimental approaches. As we conclude, we would like to draw the reader's attention to a few additional items.

First, we want to emphasize the importance of being clear on the **nature of confounding relationships**. This clarity helps determine whether the IV approach is necessary, and if necessary, helps identify a valid IV. More generally, clearly specifying the sources of confoundedness is useful first step in figuring out what methods are useful to deal with the identification challenges they present. Simpler methods, with fewer assumptions, might be sufficient and preferred. For instance, if the confounders vary at a level higher than the variation in D or Y, the researcher may be able to control for U with fixed-effects.

Second, we would like to highlight the issue of **weak instruments**. Although the inclusion restriction only requires that Z is associated with D, a large literature, which we do not cover here, shows that weak associations can mean that the 2SLS estimate is vulnerable to bias (e.g., Rossi, 2014; Andrews et al., 2019). The weak instrument bias is often exacerbated by a large number of instruments.[10] Therefore, one strong instrument is generally preferred over using many instruments, some of which are weak (Angrist and Kolesár, 2024). This should also caution readers about including a large number of fixed-effects as IVs in the model, especially if the nature of the confounder does not require them.

In order to assess the strength of the instrument, researchers often use the rule of thumb that the F-statistic (on the null hypothesis that coefficients in the first stage are zero) should be 10 or larger, even though the original research this rule-of-thumb is predicated on offers more nuanced critical values (e.g., Staiger and Stock, 1997; Stock and Yogo, 2002; Stock et al., 2002). A well known issue with using this rule of thumb is that homoskedasticity was a key assumption in the literature that produced it. In the case of one endogenous regressor and linear models, Olea and Pflueger (2013) propose an effective first-stage F-statistic that corrects for non-homoskedastic errors (e.g., clustering, auto-correlation). Lee et al. (2022) offer an inference approach for the single instrument case that is robust to heteroskedasticity and clustering, which applies an adjustment factor to the 2SLS standard errors based on the first-stage. This work shows that once violations of homoskedasticity are considered, the

[10] Intuitively, the 2SLS estimator with multiple instruments is a weighted average of the causal effects of each instrument, where the weights are related to the strength of the first-stage (Angrist and Pischke, 2009).

necessary critical values are larger by an order of magnitude compared to the common rule of thumb. We should also highlight that in the case of a weak first-stage, not all hope is lost: researchers can use weak instrument robust inference. We refer the interested reader to the econometrics literature on weak instruments (e.g., Angrist et al., 1999; Andrews and Stock, 2005; Andrews et al., 2019; Chernozhukov and Hansen, 2008) for further details.

Third, it is important for papers using the IV method to think carefully about the **external validity** of the results they obtain. As we discussed, the IV approach can only identify the treatment effect among compliers. This is an unknown subset of the data, as treated units are a mix of always-takers and compliers. Furthermore, the complier group depends on the instrument used. Different instruments will lead to different estimands. So, given the instrument(s) the researcher is using, it is important to think about the specificity of the source of variation in D that is generated by Z, and discuss how generalizable the results may be to other groups, situations, or times. Sometimes, the instrument only impacts the behavior of a narrow group of people who are likely to have different $D \to Y$ than the population of interest. At other times, we may not expect meaningful differences in the causal $D \to Y$ relationship between compliers and the general population. Research benefits from transparency. We recommend authors to openly discuss external validity issues and use institutional details to support any arguments of generalizability.

To conclude, strategic and nonrandom decisions by consumers, managers, firms, regulators, and other institutional actors permeate marketing problems and issues. For many such situations, the IV approach may be the right quasi-experimental approach to study research questions of interest. We hope this piece encourages the appropriate use of IVs as tools to provide valid theoretical and managerial insights.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Andrews, D. and J. H. Stock (2005): "Inference with Weak Instruments," *NBER working paper*.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. and M. Kolesár (2024): "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," *Journal of Econometrics*, 240.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Iinstrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. D. and A. B. Krueger (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier Science Publishers, vol. 3A, 1277–1366.

——— (2001): "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15, 69–85.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton university press.

Bruhn, J., K. Greenberg, M. Gudgeon, E. K. Rose, and Y. Shem-Tov (2022): "The Effects of Combat Deployments on Veterans' Outcomes," NBER Working paper w30622.

Burke, M., L. F. Bergquist, and E. Miguel (2019): "Sell low and buy high: arbitrage and local price effects in Kenyan markets," *The Quarterly Journal of Economics*, 134, 785–842.

Calder-Wang, S. and P. A. Gompers (2021): "And The Children Shall Lead: Gender Diversity and Performance in Venture Capital," *Journal of Financial Economics*, 142, 1–22.

Chernozhukov, V. and C. Hansen (2008): "The Reduced form: A Simple Approach to Inference with Weak Instruments," *Economics Letters*, 100, 68–71.

Chyn, E., B. Frandsen, and E. C. Leslie (2024): "Examiner and Judge Designs in Economics: A Practitioner's Guide," Tech. rep., National Bureau of Economic Research.

Cunningham, S. (2021): *Causal Inference: The Mixtape*, Yale University Press.

Danieli, O., D. Nevo, I. Walk, B. Weinstein, and D. Zeltzer (2023): "Negative controls for instrumental variable designs," *arXiv preprint arXiv:2312.15624*.

De Chaisemartin, C. (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8, 367–396.

Erikson, R. S. and L. Stoker (2011): "Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes," *American Political Science Review*, 105, 221–237.

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging Judge Fixed Effects," *American Economic Review*, 113, 253–277.

Gerber, A. S. and D. P. Green (2012): *Field Experiments: Design, Analysis, and Interpretation*, W. W. Norton.

Goldfarb, A., C. Tucker, and Y. Wang (2022): "Conducting Research in Marketing with Quasi-Experiments," *Journal of Marketing*, 86, 1–20.

Gong, S., J. Zhang, P. Zhao, and X. Jiang (2017): "Tweeting as a Marketing Tool: A Field Experiment in the TV Industry," *Journal of Marketing Research*, 54, 833–850.

Hansen, B. B. and J. Bowers (2008): "Covariate Balance in Simple, Stratified and Clustered Comparative Studies," *Statistical Science*, 219–236.

Holtz, D., R. Lobel, I. Liskovich, and S. Aral (2024): "Reducing Interference Bias in Online Marketplace Pricing Experiments," *Management Science*.

Imbens, G. W. (2020): "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *Journal of Economic Literature*, 58, 1129–1179.

Keele, L. et al. (2019): "Falsification Tests for Instrumental Variable Designs with an Application to Tendency to Operate," *Medical Care*, 57, 167–171.

Kiviet, J. F. and S. Kripfganz (2021): "Instrument Approval by the Sargan Test and Its Consequences for Coefficient Estimation," *Economics Letters*, 205, 109935.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022): "Valid t-ratio Inference for IV," *American Economic Review*, 112, 3260–3290.

Lee, N., B. Bollinger, and R. Staelin (2023): "Vertical versus Horizontal Variance in Online Reviews and their Impact on Demand," *Journal of Marketing Research*, 60, 130–154.

Li, Y. and Y. Xie (2020): "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement," *Journal of Marketing Research*, 57, 1–19.

Moshary, S., B. T. Shapiro, and J. Song (2021): "How and when to use the political cycle to identify advertising effects," *Marketing Science*, 40, 283–304.

Murray, M. P. (2006): "Avoiding Invalid Instruments and Coping with Weak Instruments," *Journal of Economic Perspectives*, 20, 111–132.

Olea, J. L. M. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Papies, D., P. Ebbes, and E. M. Feit (2023): "Endogeneity and Causal Inference in Marketing," in *The History of Marketing Science*, World Scientific, 253–300.

Pearl, J. (2009): "Causal Inference in Statistics: An Overview," *Statistics Surveys*, 3, 96–146.

Rossi, P. E. (2014): "Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications," *Marketing Science*, 33, 621–762.

Rubin, D. B. (2005): "Causal Inference using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 100, 322–331.

Sinkinson, M. and A. Starc (2019): "Ask Your Doctor? Direct-to-Consumer Advertising of Pharmaceuticals," *Review of Economic Studies*, 86, 836–881.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65, 557–586.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Stock, J. H. and M. Yogo (2002): "Testing for Weak Instruments in Linear IV Regression," Tech. rep., National Bureau of Economic Research.

Sunder, S., V. Kumar, A. Goreczny, and T. Maurer (2017): "Why do Salespeople Quit? An Empirical Examination of Own and Peer Effects on Salesperson Turnover Behavior," *Journal of Marketing Research*, 54, 381–397.

Yang, F., J. R. Zubizarreta, D. S. Small, S. Lorch, and P. R. Rosenbaum (2014): "Dissonant Conclusions When Testing the Validity of an Instrumental Variable," *American Statistician*, 68, 253–263.

Yang, J., J. Y. Lee, and P. K. Chintagunta (2021): "Commercial Success Through Commercials? Advertising and Pay-TV Operators," *Journal of Marketing Research*, 58, 925–947.

Yu, S., M. Ghosh, and M. Viswanathan (2022): "Money-Back Guarantees and Service Quality: The Marketing of In Vitro Fertilization Services," *Journal of Marketing Research*, 59, 659–673.

Table 1: Summary of Assumptions in the IV Approach

Assumption	**Critical Question**	**Calder-Wang and Gompers (2021)**	**Sinkinson and Starc (2019)**
Independence assumption: The instrument Z does not share common causes with the outcome Y	Are there any omitted variables that determine Z and Y? What allows us to claim that Z is as good as random (when X is not)?	Gender of partner's children is determined by nature, and thus independent of firm performance.	The variation in political ad spending is independent of the variation in statin demand conditions.
SUTVA: A unit's response to its own value of the instrument Z_i does not depend on the value of the instrument for other units Z_{-i}.	Is it possible that there are spillovers or interference among different units?	A partner's hiring decisions and financial performance is not impacted by the gender of other partners' children.	Advertising and sales of a pharmaceutical company in a market-month does not respond political ad spending in other markets and months.
Inclusion restriction: The Instrument Z must influence the treatment D, either in a positive or negative manner.	Why do we expect D to respond to Z? (Note: this is the only assumption for which we can provide empirical evidence)	Partners who have more daughters are more likely to hire women.	Increases in political advertising spending displace other types of advertising.
Exclusion restriction: The effect of instrument Z on outcome Y operates only through the effect of Z on D.	Can Z to influence Y through other channels (direct or indirect) that are not through D?	Gender of partners' children do not have an impact on firm performance except through its impact on hiring.	Political advertising cycle has no other effect on pharmaceutical demand except through its impact on advertising decisions.
Monotonicity: The impact of the instrument Z on D across units of analysis is (weakly) in the same direction.	Do compliers and defiers coexist?	Having more daughters would (weakly) increase the likelihood of hiring women for all partners.	All statin manufacturers (weakly) decrease advertising when political ad spending increases.

Disclaimer/Warning: The goal of this table is to summarize the discussion in the Assumptions section. It should not be used as a template or a checklist. It is not intended to support, not replace, critical engagement with the necessary identifying assumptions in a given empirical context.