*American Economic Review 2026, 116(4): 1380–1414*
*https://doi.org/10.1257/aer.20240636*



# Negative Control Falsification Tests for Instrumental Variable Designs[†]

*By* Oren Danieli, Daniel Nevo, Itai Walk, Bar Weinstein, and Dan Zeltzer*

*The validity of instrumental variable (IV) designs is typically tested using two types of falsification tests. We characterize these tests as conditional independence tests between negative control variables—proxies for unobserved variables posing a threat to the identification—and the IV or the outcome. We describe the conditions variables must satisfy in order to serve as negative controls. We show these falsification tests examine not only independence and the exclusion restriction but also functional form assumptions. Our analysis reveals conventional applications of these tests may flag problems even in valid IV designs. We offer implementation guidance to address these issues.* (*JEL* C12, C18, C26, C52)

When examining the identification assumptions in instrumental variable (IV) designs, researchers often use indirect falsification, or "placebo," tests. Reviewing the most-cited papers in five leading economics journals, we find that 51 percent of IV studies employ such falsification tests. The large majority of falsification tests fall into two categories: 75 percent of papers that implemented a falsification test examined that the IV is not associated with certain variables, such as lagged outcomes, which we call *negative control outcomes* (NCOs). Similarly, 24 percent examined that the outcome is not associated with other variables, which we call *negative control instruments* (NCIs). For example, they tested that the outcome is not associated with variables that resemble the IV but do not affect the treatment. While negative controls have been extensively studied in classic causal settings (Lipsitch, Tchetgen Tchetgen, and Cohen 2010; Shi, Miao, and Tchetgen Tchetgen 2020),

*Danieli: Tel Aviv University (email: orendanieli@tauex.tau.ac.il); Nevo: Tel Aviv University (email: danielnevo@tauex.tau.ac.il); Walk: Tel Aviv University (email: itaiwalk@mail.tau.ac.il); Weinstein: Tel Aviv University (email: barwein@mail.tau.ac.il); Zeltzer: Tel Aviv University and University of California, Berkeley (email: dzeltzer@berkeley.edu). Andres Santos was the coeditor for this article. We thank Maya Orenstein for excellent research assistance. We thank Kirill Borusyak, Yoav Goldstein, Matan Kolerman, David Lee, Ro'ee Levy, Ashesh Rambachan, Tamir Zehavi, and seminar participants at the ASSA 2023 Meeting, EUROCIM 2024, Princeton University, Stanford University, UCSB, UCSC, Technion, Hebrew University of Jerusalem, Tel Aviv University, EIEF Conference, SOLE 2024 meeting, American Health Econometrics Workshop, and EWMES 2025 for their helpful comments and suggestions. We acknowledge support from the Tel Aviv University Center for AI and Data Science (TAD) in collaboration with Google as part of the initiative of AI and DS for social good and from the Sapir Center for Development.

[†]Go to https://doi.org/10.1257/aer.20240636 to visit the article page for additional materials and author disclosure statement(s).

their use in falsification tests for the assumptions underlying IV designs remains underexplored.[1] This paper aims to fill this gap.

We propose that practitioners first identify potential threats to their IV validity, which we define formally. Since variables posing these threats are typically unobserved, researchers should look for proxy variables, termed *negative controls*. We characterize the conditions that these proxies must satisfy and show how to use them to test the validity of the IV design using (conditional) independence tests. We highlight two common pitfalls that can falsely flag problems in valid IV designs. First, NCI tests typically require conditioning on the IV—a step frequently overlooked in practice. Second, prevalent negative control tests may flag violations of unnecessary or replaceable functional form assumptions. We propose ways to separately test the validity of the IV design from these functional form assumptions. Our framework also suggests novel, underutilized negative control variables and testing methods.

We first introduce the concept of an *alternative path variable*, which represents a variable that poses a threat to the identification. In valid IV designs, which satisfy independence and the exclusion restriction, the only path between the IV and the outcome is through the treatment.[2] Threats to identification can be characterized as alternative paths between the IV and the outcome through an alternative path variable, rather than through the treatment.

To construct a negative control test, researchers need to determine which type of alternative path variable threatens the IV design. We distinguish between two categories of variables that can create such paths. In the first category, alternative path outcome (APO) variables, the concern is that a variable that is associated with the outcome would also be associated with the IV. Figure 1 illustrates two examples of such cases.[3] Panel A shows a potential violation of the independence assumption. For concreteness, consider the context of Martin and Yurukoglu (2017), who examine the impact of Fox News viewership ($X$) on Republican vote shares ($Y$). As an IV, they use the local Fox News cable channel position ($Z$) since lower channel numbers induce greater viewership. One concern is that unobserved local conservativeness ($U_1$, the APO variable) affects not only the Republican vote share (the outcome) but also the channel position (the IV), as marked by the dashed arrow. If that is the case, an alternative path between the channel position and Republican voting share exists via local conservativeness, the APO variable. This would violate the independence assumption and invalidate the design. Panel B offers an example of an APO variable ($U_2$) that is part of a potential violation of the exclusion restriction assumption.

In the second category, alternative path instrument (API) variables, the concern is that variables known to be associated with the IV would also be associated with the outcome. Panel C of Figure 1 provides an example of an API variable that potentially violates the independence assumption. For concreteness, consider the context of Nunn and Qian (2014a), who examine the effect of US food aid ($X$) on conflicts

[1] In epidemiology and biomedical fields, researchers use the similar terminology of negative controls for falsification tests to detect potential confounding in an exposure-outcome relationship.

[2] The independence assumption typically includes independence of the IV with the potential treatment and the potential outcome (Abadie 2003). The falsification tests we discuss in this paper focus only on outcome independence. See Section IIA.

[3] We use directed acyclic graphs (DAGs) to visualize complex structures, as advocated by Imbens (2020). Formal results using the causal DAG framework (Pearl 2009b) appear in Supplemental Appendix B.

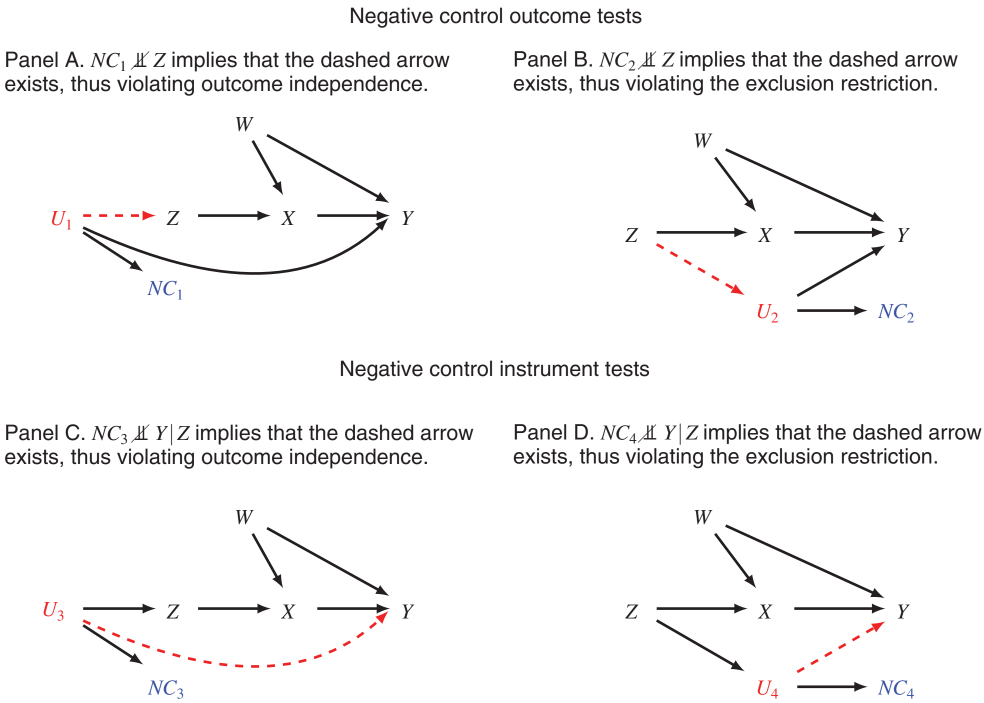


Figure 1. Negative Control Falsification Tests: Graphical Illustrations

*Notes*: The figure illustrates how negative control tests assess the validity of IV designs. In all the panels, $X$ represents the endogenous treatment variable, $Y$ the outcome, $Z$ the IV, and $W$ a potential confounder that motivates the use of IV. The variables $U_i$ are unobserved variables that threaten identification when the dashed arrows exist. The top panels (A and B) depict negative control outcome tests. IV validity is threatened by the concern that $U_1$ or $U_2$ (*alternative path outcome* (APO) variables) are related to the IV, thus violating outcome independence (panel A) or the exclusion restriction (panel B). An observed negative control outcome ($NC_1$ or $NC_2$) related to each APO variable can be used to evaluate the presence of the problematic association by testing whether $NC_i \perp\!\!\!\perp Z$. The bottom panels (C and D) depict negative control instrument tests, addressing concerns that $U_3$ or $U_4$ (*alternative path instrument* (API) variables) might be related to the outcome, thus violating outcome independence (panel C) or the exclusion restriction (panel D). An observed negative control instrument ($NC_3$ or $NC_4$) related to each API variable can be used to examine these concerns by testing whether $NC_i \perp\!\!\!\perp Y|Z$.

in recipient countries ($Y$). They use the US production of wheat ($Z$), a staple aid crop, as an IV for aid. Here, the API variable is unobserved weather conditions ($U_3$). It is known that weather conditions affect wheat production, the IV. The question is whether they also affect conflicts, the outcome, as marked by the dashed arrow. If so, an alternative path from wheat production to conflict exists via the API variable. Panel D offers an example of an API variable ($U_4$) that is part of a potential violation of the exclusion restriction assumption.

Since alternative path variables of both categories are often unobserved, researchers can utilize negative control variables as proxies for them. A negative control outcome (NCO) is a proxy for an APO variable. An association between an NCO and the IV implies the presence of an alternative path, indicating that the design is not valid. In Martin and Yurukoglu (2017), the lagged outcome, Republican vote share in 1996 ($NC_1$ in Figure 1) is used as an NCO for unobserved conservativeness ($U_1$). If 1996 vote shares is associated with channel position, it would imply that

Table 1—Examples of Negative Control Tests for IV Designs in Economics

| Paper | Treatment | Outcome | IV | Threat | NCO (IV should not be correlated with it) |
|---|---|---|---|---|---|
| *A. Negative control outcome tests* | | | | | |
| Martin and Yurukoglu (2017) | Fox News viewership | Republican vote share in 2008 | Channel position: Lower channel numbers induce larger viewership | Cable companies might place Fox News in lower channels in more conservative locations | Republican vote share in 1996 |
| Angrist and Evans (1998) | Number of children | Female labor supply | Same-sex sibship: Families with same-sex sibship for the first two children are more likely to have more children | Same-sex sibship may increase hand-me-downs, reducing expenditures and potentially labor supply | Clothing expenditure (Rosenzweig and Wolpin 2000) |
| Autor, Dorn, and Hanson (2013a,b) | Shift-share based on import penetration in US | Manufacturing employment | Shift-share based on import penetration in non-US developed countries | Commuting zones with importing industries might be declining for other reasons | Manufacturing employment trends before large Chinese import competition |
| Doyle et al. (2015); Chan, Card, and Taylor (2023) | Hospital assignment | Health outcomes | Ambulance company assignment, which strongly predicts hospital assignment | Patient ambulance assignment may depend on their health | Patient demographics |
| Jäger, Heining, and Lazarus (2024) | Worker exits | Incumbent workers' wages | Unexpected worker deaths in small firms | Worker death could be correlated with firm-specific risks, which affect wages | Worker unexpected deaths in same firms in other years |

(*continued*)

cable companies consider the population conservativeness when placing the channels, such that the dashed arrow exists. Hence, there is an alternative path between the IV and the outcome, violating the independence assumption. Panel A of Table 1 describes further examples of applications of NCO variables in economic research.

Similarly, a negative control instrument (NCI) can be used as a proxy for an API variable. For example, Nunn and Qian use orange production as an NCI ($NC_3$ in Figure 1) for unobserved weather conditions $(U_3)$; orange production is affected by similar weather conditions as wheat but is not used for food aid. If orange production were associated with conflicts, conditional on wheat production, this would indicate an alternative path between the IV and the outcome, violating the independence assumption. Panel B of Table 1 lists additional examples of NCI variables in economic research.

The definition of negative controls clarifies which proxy variables researchers can use as NCOs or NCIs. This definition guarantees that these proxies can test

Table 1—Examples of Negative Control Tests for IV Designs in Economics (*continued*)

| Paper | Treatment | Outcome | IV | Threat | NCI (conditional on IV, outcome should not be correlated with it) |
|---|---|---|---|---|---|
| *B. Negative control instrument tests* | | | | | |
| Nunn and Qian (2014a) | US food aid | Conflict in recipient countries | Wheat production: US food aid increases when it booms | Wheat production is affected by weather conditions, which could also have other impacts on conflicts | Production of crops not used for aid (e.g., oranges) |
| Ashraf and Galor (2013a) | Genetic diversity | Economic development | Distance from Addis Ababa, which predicts genetic diversity due to the human origin hypothesis | Other economic factors have geographical dispersion | Distance from other cities (e.g., London, Mexico City) |
| De Giorgi, Frederiksen, and Pistaferri (2020) | Peer consumption | Own consumption | Shocks in firms of distant peers: Negative shocks in firms of distant peers are less likely to be correlated with self economic shocks | Shocks to larger firms are statistically more likely to affect distant peers, and might also affect consumption in other ways | Placebo shocks: Calculate same IV based on permuted employer-employee relationship (keeping employer size unchanged) |
| Madestam et al. (2013) | Tea Party protest participation | Republican vote | Rain on April 15, 2009, which affected local participation in one of the first large Tea Party protests | Probability of rain is driven by local climate conditions, which could relate to voting in various ways | Rain on other dates |

for alternative paths using (conditional) independence tests. In particular, variables directly associated with the IV (not through the APO variable) cannot serve as NCOs. Similarly, variables associated directly with the outcome (not through the API variable or the IV) cannot serve as NCIs. Tests using such variables would falsely flag valid IV designs.

We highlight two common pitfalls in current practice. First, in our survey, the vast majority of papers using NCI variables implemented a test that will falsely flag problems in valid IV designs (with sufficient sample size). In almost all these papers, the NCI was a variable that is similar to the IV but does not affect the treatment (such as orange instead of wheat production in the previous example). Researchers typically test whether such an NCI is associated with the outcome by plugging it instead of the original IV in the reduced-form equation. This specification overlooks a key issue: NCIs are typically associated with the original IV and therefore will be associated with the outcome, even in a valid IV design. For example, as shown in panel C of Figure 1, both orange production ($NC_3$) and wheat production ($Z$) are influenced by weather conditions ($U_3$). Hence, orange production and conflict ($Y$)

would be dependent even if wheat production is a valid IV (i.e., the dashed arrow does not exist, and there is no alternative path). We show that this problem can be avoided by controlling for the original IV in the NCI test. In the above example, this means controlling for wheat production when testing the association between orange production and conflict.

The second pitfall is that negative control tests may flag problems even in valid IV designs due to a misspecified functional form. In 2SLS specifications, researchers must choose a functional form for the IV and the control variables, typically assuming a simple linear-additive model. This structure often carries over into the execution of negative control tests. Consequently, even valid IV designs may fail negative control tests merely due to violations of functional form assumptions. However, unlike the independence and exclusion assumptions, which are necessary for IV validity, functional form assumptions can often be relaxed and, in some cases, are unnecessary for identifying causal effects. To address this problem, researchers can use alternative negative control tests that rely on weaker functional form assumptions.

The framework can also be used to identify new types of negative control variables, some of which have yet to be commonly employed in empirical research. For example, variables that causally influence the IV could serve as NCIs (e.g., observed weather conditions can serve as NCIs when the IV is wheat production).

This paper adds to prior econometrics work on tests for IV design validity and, more generally, the validity of causal designs. Recent work has suggested novel tests to examine the validity of IV designs (Kitagawa 2015; Huber and Mellace 2015; Mourifié and Wan 2017; Frandsen, Lefgren, and Leslie 2023; Chyn, Frandsen, and Leslie 2025). Previous work has also discussed robustness tests, not specific to IV, based on varying the set of controls (Altonji, Elder, and Taber 2005; Oster 2019; Diegert, Masten, and Poirier 2022). However, Pei, Pischke, and Schwandt (2019) recommend using such control variables as NCOs instead. Eggers, Tuñón, and Dafoe (2024) discuss the usage of placebo tests in the social sciences more broadly. We contribute to this literature by studying the most common type of falsification tests for IV designs.

This paper also contributes to the growing literature on negative controls (Lipsitch, Tchetgen Tchetgen, and Cohen 2010; Shi, Miao, and Tchetgen Tchetgen 2020). In a standard causal design, negative controls are used to detect or even correct for bias due to unobserved treatment-outcome confounding. To this end, valid and, in some cases, even invalid IVs can serve as *negative control exposures* and under further assumptions can be combined with an additional negative control to achieve point identification (Miao, Geng, and Tchetgen Tchetgen 2018; Shi, Miao, and Tchetgen Tchetgen 2020; Tchetgen Tchetgen et al. 2024; Dukes et al. 2025). We study negative controls in a different setting, where researchers employ an IV design and seek to use negative controls specifically for testing the IV assumptions in order to assess the design's validity. We find several important differences in the application of negative controls for assessing IV designs compared to their usage in the standard treatment-outcome setting. Our work is also related to Davies et al. (2017), who use negative controls in IV contexts without developing a general framework for such an approach.

Finally, negative control tests can also be derived from the causal DAG framework (Pearl 2009a,b). In particular, negative control tests can be seen as an application

Table 2—Current Use of Falsification Tests for IV Designs in Economics

| | | | Falsification test characteristics (Share of papers that included falsification tests) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Type of test | | | Test specification | | | |
| | Papers reviewed (1) | Share with falsification tests (2) | Negative control outcome (3) | Negative control instrument (4) | Other (5) | Replace outcome with NCO (6) | Replace IV with NCI (7) | Add NCI to reduced form (8) | No. negative controls used (medium) (9) |
| All | 140 | 0.51 | 0.75 | 0.24 | 0.21 | 0.43 | 0.18 | 0.06 | 3.50 |
| *By journal* | | | | | | | | | |
| REStud | 48 | 0.42 | 0.75 | 0.40 | 0.10 | 0.65 | 0.35 | 0.05 | 4.00 |
| AER | 42 | 0.50 | 0.86 | 0.24 | 0.05 | 0.57 | 0.24 | 0.00 | 4.00 |
| JPE | 21 | 0.62 | 0.62 | 0.15 | 0.31 | 0.23 | 0.00 | 0.15 | 3.50 |
| QJE | 19 | 0.68 | 0.77 | 0.15 | 0.46 | 0.15 | 0.08 | 0.08 | 2.00 |
| ECMA | 10 | 0.50 | 0.60 | 0.00 | 0.40 | 0.20 | 0.00 | 0.00 | 4.50 |

*Notes:* The table shows the results of our survey of highly cited articles employing instrumental variable designs published in leading economics journals from 2013 to 2023. The sample includes all articles from this period in the *Review of Economic Studies* (REStud), *American Economic Review* (AER), *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE), and *Econometrica* (ECMA) that used IV designs and had significant citation counts on Google Scholar (over 300 citations for papers until 2020 and over 100 for those published after 2020). We examined these papers for their use of falsification tests. Column 2 shows the proportion of papers that employ any falsification test. Columns 3–8 report the fraction of papers that implemented different types of falsification tests out of all papers that implemented any falsification test. Columns 3–5 categorize the tests into negative control outcome, negative control instrument, and other types of falsification tests, respectively. The fractions do not sum up to one, as some papers employed multiple test types. Columns 6–8 provide the share of papers using simple linear designs where the reduced-form equation is modified by replacing the outcome with an NCO, replacing the IV with an NCI, and adding the NCI (while still conditioning on the original IV). Column 9 reports the median number of negative control variables used. Supplemental Appendix A provides additional details on the survey construction.

of the broader principle that a fully specified causal DAG yields testable implications in the observed data. In fact, such implications can be derived algorithmically once the DAG is specified (Textor et al. 2016). Hence, when researchers are able to articulate a complete DAG representing their design, they can use these results to generate falsification tests in their setting.[4]

The rest of this paper proceeds as follows. Section I surveys the current practice of falsification tests for IV designs. Section II presents a potential-outcomes framework for negative control tests in IV designs. Section III provides guidance for practitioners and demonstrates key findings using recent empirical studies. Section IV concludes.

## I. Survey of Current Practice

To provide an overview of current practices in falsification testing for IV designs, we surveyed the most highly cited articles with an IV analysis published between 2013 and 2023 in top economics journals. We then classified the characteristics of the falsification tests used. Supplemental Appendix A provides additional details on the survey construction, and the results are summarized in Table 2.

[4] In Supplemental Appendix B, we apply these principles to show explicitly how DAG theory can be used for constructing falsification tests for IV designs.

We highlight five key findings from this survey. First, falsification tests are widely used in IV analyses. Approximately half (51 percent) of all articles surveyed employ some form of falsification test (column 2 of Table 2).

Second, most falsification tests fall within the negative control framework we study in this paper. As outlined earlier, these tests can be divided into two types: *negative control outcome* tests, which check for associations between the IV and variables it should not be associated with, and *negative control instrument* tests, which examine associations between the outcome and variables it should not be associated with. Among surveyed papers using falsification tests, 75 percent used NCO tests (column 3), and 24 percent used NCI tests (column 4). All other types of falsification tests combined were used in 21 percent of the papers (column 5); Supplemental Appendix A lists these other, less common types.

Third, current applied work usually restricts itself to two simple types of negative control test specifications that rely on the 2SLS functional form assumptions. NCO tests typically involve estimating a revised reduced-form equation using an alternative outcome (e.g., a lagged outcome) and testing if it is unrelated to the IV (column 6). Such specifications account for 57 percent of all NCO tests. The remaining NCO tests often follow a similar logic.[5] For NCI tests, researchers either replace the original IV in the reduced-form equation with a similar variable that does not affect the treatment (column 7), or add this variable to the reduced-form equation (column 8).

Fourth, most reported NCI tests are implemented incorrectly. As noted in the introduction and discussed in detail later, NCI tests should almost always control for the original IV. In practice, only 24 percent of the papers using NCI tests reported doing so. With sufficient sample size, this error leads to finding false problems in valid IV designs. It is likely that additional NCI tests were conducted incorrectly, finding false problems in valid designs, and were therefore not reported.

Finally, papers using falsification tests usually utilize only a few negative control variables. The median number of negative control variables used in the surveyed papers is 3.5 (column 9), with 35 percent of the papers using only 1. This finding suggests that researchers use only a subset of the available negative controls. As we demonstrate in Section III, our framework can guide a systematic search for negative control variables in existing data and suggest novel types of negative control variables researchers can use to evaluate their IV designs.

## II. Formalizing Negative Controls in IV Settings

We use the terminology of potential outcomes to formalize negative control tests for IV designs; we use DAGs for examples and intuition. Supplemental Appendix B introduces the basic relevant concepts for DAG theory and presents how the key results presented in this paper can be derived using DAGs.

[5] For example, balance tables that regress various NCOs on the IV.

### A. *The Threats to the Identification*

*Assumptions*.—Consider i.i.d. units indexed by $i = 1, \ldots, n$. Denote the observed (endogenous) treatment status by $X_i$, and the candidate IV by $Z_i$. Let $Y_i(z,x)$ be the potential outcome for unit $i$ had $Z_i$ and $X_i$ been jointly set to the values $z$ and $x$, respectively.[6] We make the standard assumption that the observed outcome $Y_i$ is given by $Y_i = Y_i(Z_i,X_i)$. Because units are assumed to be i.i.d., we omit the subscript $i$ when it improves clarity. All variables may be discrete or continuous.

The negative control tests we discuss in this paper examine whether there is an alternative path between the IV and the outcome, in addition to the standard path through the treatment. Such an additional path would violate one of the following two assumptions. The first assumption, *outcome independence*, maintains that IV assignment is independent of the potential outcomes.

ASSUMPTION 1 (Outcome Independence): *For all $z$ and $x$*, $Z \perp\!\!\!\perp Y(z,x)$.

This assumption is usually written as part of a broader independence assumption (e.g., Abadie 2003). We distinguish between outcome independence and *treatment independence*, which requires $Z \perp\!\!\!\perp X(z)$ for every value of $z$. Only outcome independence is tested in the negative control tests we discuss in this paper.

Outcome independence is violated if the IV is affected by a variable that also affects the outcome. As previously discussed, Martin and Yurukoglu (2017) study the effect of Fox News viewership on voting using channel positions as an IV. This example is illustrated in panel A of Figure 1. The concern is that cable companies accounted for local conservativeness $(U_1)$ when assigning Fox News channel position $(Z)$, as illustrated by the dashed arrow. If so, an alternative path emerges between channel position and voting $(Y)$ through local conservativeness. This path violates outcome independence.

The second assumption, *exclusion restriction*, maintains that the IV does not have a direct effect on the outcome. Let $Y(x)$ be the potential outcome had the treatment $X$ been set to $x$, while $Z$ had not been set to any particular value and takes its natural value, that is, $Y(x) = Y(Z,x)$.

ASSUMPTION 2 (Exclusion Restriction): *For all $z,x$*, $Y(z,x) = Y(x)$.[7]

The exclusion restriction is violated if the IV affects the outcome in alternative ways, in addition to its effect through the treatment.

Panel B of Figure 1 illustrates a potential violation of the exclusion restriction assumption. For example, Angrist and Evans (1998) study the effect of the number of children $(X)$ on female labor supply $(Y)$, using the sex composition of the first two children as an IV $(Z)$ (because same-sex children induce further births for parents with a preference for gender variety). One potential concern is that same-sex sibship could reduce housing expenditures $(U_2)$ due to hand-me-downs, which could then affect female labor supply decisions. In this case, an alternative path between

[6] This formulation implicitly assumes the stable unit treatment value assumption (SUTVA).

[7] Equality between random variables denotes equality almost surely (e.g., $\Pr(Y(z,x) = Y(x)) = 1$).

the sex composition of the first two children and female labor supply would exist through the effect on household expenditures. This violates the exclusion restriction.

Together, outcome independence and exclusion restriction imply[8]

$$Z \perp\!\!\!\perp Y(x) \text{ for all } x. \tag{1}$$

In loose terms, (1) requires that there are no alternative paths between the IV and the outcome except through the treatment. Because potential outcomes are never observed, neither these two assumptions nor (1) can be tested directly.

To identify a causal effect using an IV design, researchers often impose additional assumptions. For example, the design of Angrist, Imbens, and Rubin (1996) also requires treatment independence, relevance, and monotonicity. However, the negative control tests presented in this paper do not test these other assumptions.

*Alternative Path Variables*.—To formalize the notion of a threat to the identification, we introduce the concept of an *alternative path variable*—a variable that is part of a suspected alternative path between the IV and the outcome that, if such a path exists, would violate outcome independence or exclusion. For simplicity, we assume that only one potential threat to the IV validity exists. Supplemental Appendix C addresses a more general case with multiple threats. We distinguish between two types of alternative path variables that require different types of falsification tests.

The first type of identification threat involves an *alternative path outcome* variable. This variable is presumably associated with the outcome. The threat is that it is also associated with the IV, which would generate an alternative path. In Martin and Yurukoglu (2017), conservativeness of the local population is an APO variable ($U_1$ in panel A of Figure 1). Since conservativeness certainly affects voting behavior, it poses a threat if it also affects cable companies' decisions for channel position (represented by the dashed arrow). Panel B of Figure 1 describes an APO variable for a potential violation of the exclusion restriction.

The formal definition of an APO variable is as follows.

DEFINITION 1 (Alternative Path Outcome Variable): *A random variable U is an APO variable if the following two conditions hold.*

(*i*) *Latent IV validity.* $Z \perp\!\!\!\perp Y(x) \mid U$.

(*ii*) *Path indication. If* $Z \perp\!\!\!\perp Y(x)$, *then* $Z \perp\!\!\!\perp U$.

Latent IV validity posits that had we observed and conditioned on the APO variable, the IV design would have been valid (both outcome independence and the exclusion restriction would hold conditional on the APO variable). This condition implies that imperfect proxies for variables posing identification threats cannot themselves be APO variables, as controlling for an imperfect proxy does not

[8] When the exclusion restriction does not hold, $Y(x)$ is still well defined but differs from $Y(z,x)$ for some $z$. Therefore, since $Y(x) = Y(Z,x)$, violation of the exclusion restriction implies $Z \not\perp\!\!\!\perp Y(x)$.

make the IV and the potential outcome conditionally independent. In the example of Martin and Yurukoglu (2017), the share of Republican votes in 1996 is only a proxy for the APO variable (latent conservativeness), and hence, controlling for it does not eliminate the threat. Therefore, the Republican vote share in 1996 is not an APO variable, as it does not satisfy the latent IV validity condition. Latent IV validity is analogous to the latent exchangeability assumption appearing in recent literature on negative controls in epidemiology (Shi, Miao, and Tchetgen Tchetgen 2020) and statistics (Tchetgen Tchetgen et al. 2024).

Path indication states that a valid IV is not associated with the APO variable. Its contrapositive ensures that an association between the IV and the APO variable implies an alternative path between the IV and the potential outcome. Path indication guarantees that if there is a path from the IV to the APO variable, the path continues from the APO variable to the outcome. Therefore, it excludes variables unrelated to the outcome, as they can be associated with the IV without implying anything about the design validity. While APO variables often causally affect the outcome, it is not mandatory (as demonstrated in Supplemental Appendix D.1). Path indication also rules out variables that could be related to both the IV and the outcome without generating an association between them. For example, this would occur if a variable is associated with the outcome for some subpopulation but potentially associated with the IV only for a separate subpopulation. Two examples are provided in Supplemental Appendix D.2 and Supplemental Appendix D.3.

The second type of identification threat involves an *alternative path instrument* variable. An API variable is known to be associated with the IV, and the concern is that it might also be associated with the outcome. This is in contrast to an APO variable, which is known to be associated with the outcome, and the concern is its possible association with the IV. Using the previous example of Nunn and Qian (2014a), illustrated in panel C of Figure 1, weather conditions $(U_3)$ is an API variable. Wheat production $(Z)$ is known to be affected by weather. An alternative path that threatens identification may form if weather also affects conflicts $(Y)$ directly (the dashed arrow exists). Panel D describes an API variable for a potential violation of the exclusion restriction.

Formally, an API variable satisfies the following definition.

DEFINITION 2 (Alternative Path Instrument Variable): *A random variable U is an API variable if the following two conditions hold.*

$(i)$ *Latent IV validity.* $Z \perp\!\!\!\perp Y(x) \,\big|\, U.$

$(ii)$ *Path indication. If* $Z \perp\!\!\!\perp Y(x)$, *then* $U \perp\!\!\!\perp Y | Z.$

This definition resembles the definition of APO variables (Definition 1). The first condition, latent IV validity, is exactly as before. The difference between API and APO variables is encapsulated in the second condition, path indication. For API variables, this condition requires that if the IV design is valid (so $Z \perp\!\!\!\perp Y(x)$), then the API variable must be independent of the observed outcome conditional on the IV (i.e., $U \perp\!\!\!\perp Y|Z$). Through its contrapositive, this condition implies that an association between the API variable and the outcome, not via the IV, indicates that there exists an alternative path between the IV and the outcome. Therefore, the IV is invalid. Typically, path indication is satisfied when the API variable is associated with the IV.

For API variables, path indication rules out variables that are associated with the outcome through the treatment (conditional on the IV). Such variables are not informative about the validity of the IV design, as they are associated with the outcome through the treatment, even if the IV design is valid. This is different from APO variables that could be associated with the treatment (even conditional on the IV). See Supplemental Appendix D.4 for an example and a further discussion of this issue. This implies that APO variables can be associated with (or even identical to) the original confounder of the treatment-outcome relationship (which we mark by $W$ in our examples). However, an API variable cannot.

### B. *Negative Control Variables*

Negative control variables are observed proxies for the unobserved alternative path variables. Building on the definitions of alternative path variables, we are now ready to formalize the assumptions required for a random variable to serve as a negative control. The first type of negative control, *negative control outcome*, is a proxy for an APO variable. Observed variables can serve as NCOs if they satisfy the following definition.

DEFINITION 3 (Negative Control Outcome): *A random variable NC is an NCO if there exists an APO variable U such that the following two conditions hold.*

(*i*) *The NCO assumption.* $NC \perp\!\!\!\perp Z \mid U$.

(*ii*) *U-comparability.* $NC \not\perp\!\!\!\perp U$.

The NCO assumption implies that any association between the IV and the NCO must be driven by their associations with the APO variable $U$. Using DAG terminology, the NCO assumption generally implies that any open path between the IV and the NCO must go through the APO variable $U$. It rules out variables with other open paths to the IV. Panel A of Figure 1 demonstrates the NCO assumption in a setting with a potential violation of outcome independence. For example, consider again the effect of Fox News on voting (Martin and Yurukoglu 2017). The lagged outcome, Republican vote share in 1996, is an NCO ($NC_1$). The NCO assumption requires that any association between lagged voting and later channel-position ($Z$) arises only due to local conservativeness ($U_1$, the APO). Panel B of Figure 1 depicts an example of an NCO ($NC_2$) that satisfies this assumption, where a violation of the exclusion restriction is the concern.

The NCO assumption is violated for variables that are directly related to the IV, not through an APO variable. This can occur if the IV affects the candidate variable for NCO, either directly or through the treatment or the outcome. For example, various IV studies on the impacts of exposure to air pollution on different outcomes use the number of nonrespiratory hospital admissions, a seemingly unrelated outcome, as an NCO. One might expect that these admissions would be associated only with flawed IVs for air pollution. However, Guidetti, Pereda, and Severnini (2021) demonstrate otherwise. They find that air pollution increases nonrespiratory admissions through hospital congestion caused by a surge in respiratory admissions. Therefore, nonrespiratory admissions are not informative about the IV validity, as they are associated with both flawed and valid IVs. Formally, an association between nonrespiratory admissions and the IV is generated not only through an APO variable but also due to the unrelated mechanism of congestion. Hence, nonrespiratory admissions violate the NCO assumption.

The $U$-comparability assumption guarantees that the NCO has a path to the APO variable. This assumption guarantees that the NCO is a relevant proxy for the APO variable. For example, voting in 1996 satisfied $U$-comparability, as it is associated with the APO, unobserved conservativeness in the region. This assumption rules out variables that are uninformative about the design validity because they are unrelated to the identification threat (and specifically to the APO variable).

The second type of negative control variable, negative control instrument, is a proxy for API variables. Observed variables can serve as NCIs if they satisfy the following definition.

DEFINITION 4 (Negative Control Instrument): *A random variable NC is an NCI if there exists an API variable U such that the following two conditions hold.*

(*i*) *The NCI assumption.* $NC \perp\!\!\!\perp Y | Z, U$.

(*ii*) *U-comparability* $NC \not\perp\!\!\!\perp U | Z$.

The NCI assumption implies that, conditionally on the IV ($Z$), any association between the outcome and the NCI is driven by their (conditional) association with the API variable $U$. In DAG terminology, the NCI assumption generally implies that controlling for $Z$, any open path between the outcome and the NCI must go through the API variable $U$. It rules out variables that have other open paths to the outcome.

While similar, the NCI assumption and the NCO assumption (Definition 3) differ in three key aspects. First, the alternative path variable $U$ is an API variable instead of an APO variable. Second, the conditional independence is between the NCI and the outcome instead of the IV. Due to these two differences, the NCI tests defined below test for a potential association with the outcome and not with the IV. The third difference is that the statistical independence requirement is also conditional on the IV. This is because in valid IV designs, the NCI is often associated with the outcome through the IV, as we discuss in the next section.

Panel C of Figure 1 demonstrates the NCI assumption in a setting with a potential violation of outcome independence. For example, consider again the context of Nunn and Qian (2014a), which uses an alternative crop (e.g., oranges) production as an NCI ($NC_3$). Orange production is affected by similar weather conditions ($U_3$) as wheat ($Z$) and, therefore, would be associated with it. However, unlike wheat, oranges are not used as food aid ($X$). Therefore, orange production is unrelated to conflicts ($Y$), conditional on both weather and wheat production.

In many applications, the NCI assumption rules out a large class of observed variables because of their association with the outcome. For example, demographic variables often exhibit an association with the outcome, even conditional on the IV and the API variable, and therefore cannot serve as NCIs. Variables that are associated with the treatment are also not NCIs, as they are also associated with the outcome conditional on the IV and the API variable. In practice, the NCI assumption is more restrictive than the NCO assumption. The reason is that the NCO assumption requires conditional independence with the IV, which is typically more plausible than conditional independence with the outcome.

The *U*-comparability assumption for NCIs implies that the NCI is indeed a proxy for the API variable. In contrast to *U*-comparability for NCOs, for NCIs, their association with the API variable must exist conditionally on the IV. This assumption rules out variables that are not informative about the IV validity because they are unrelated to the identification threat (and specifically to the API variable), conditional on the IV.

The NCO and NCI definitions are analogous to the conditions that were formalized in previous literature on negative controls. In particular, U-comparability is common in the literature on negative controls (Lipsitch, Tchetgen Tchetgen, and Cohen 2010; Shi, Miao, and Tchetgen Tchetgen 2020). The NCO and NCI assumptions are similar to the conditional independence assumption of Tchetgen Tchetgen et al. (2024) (see equations (12) and (13)).

Definitions 3 and 4 imply that alternative path variables are themselves negative controls. This is because APO and API variables trivially satisfy both conditions in the definitions. For example, in the previously discussed design of Angrist and Evans (1998), the concern is that the sex composition of the first two children (the IV) may influence household expenditures (the APO variable) due to hand-me-downs, forming an alternative path to female labor supply (the outcome). Rosenzweig and Wolpin (2000) explore this by using a dataset in which clothing expenditures are observed and use this APO variable as an NCO.

The NCO and NCI assumptions can be weakened to cover more variables that are informative about the validity of the IV design. In Supplemental Appendix C, we offer a more general definition of negative controls that allows for direct associations between the NCO and the IV or the NCI and the outcome, not through the alternative path variable if the design is invalid.

## C. *Negative Control Tests*

A *negative control outcome test* (NCO test) is any statistical test of independence between the IV and an NCO. The null hypothesis is $H_0 : Z \perp\!\!\!\perp NC$. For example, Martin and Yurukoglu (2017) regress their NCO, Republican vote share in 1996, on the IV, Fox channel positioning. Under the null, the coefficient on the IV in this regression model should equal zero. Indeed, they were unable to reject a zero coefficient.

The following theorem states that rejecting the null hypothesis implies a violation of outcome independence or the exclusion restriction.

THEOREM 1: *Assume that a random variable NC is an NCO* (*Definition 3*). *If* $NC \not\!\perp\!\!\!\perp Z$, *then either outcome independence or exclusion restriction is violated. That is*, *the IV design is invalid.*

All proofs are given in Supplemental Appendix C. The Supplemental Appendix proof covers a more general version of Theorem 1 for designs that include control variables (discussed in Section IID). For the case without controls, the sketch of the proof is as follows. By the NCO assumption, the dependence between the IV and the NCO implies an association between the IV and an APO variable ($Z \not\!\perp\!\!\!\perp U$). By path indication, $Z \not\!\perp\!\!\!\perp U$ indicates an alternative path between the IV and the outcome ($Z \not\!\perp\!\!\!\perp Y(x)$); that is, the IV design is invalid.

Similarly, a *negative control instrument test* (NCI test) examines whether the outcome and the NCI are independent, conditional on the IV. Formally, the statistical test is for the null hypothesis $H_0 : NC \perp\!\!\!\perp Y|Z$. If the NCI is associated with the outcome conditional on the IV, this necessarily implies that the IV design is not valid, as stated in the following theorem.

THEOREM 2: *Assume that a random variable NC is an NCI (Definition 4). If $NC \not\!\perp\!\!\!\perp Y|Z$, then either outcome independence or exclusion restriction is violated. That is, the IV design is invalid.*

For example, Nunn and Qian (2014a) regress their outcome, conflicts, on various alternative crop production (e.g., oranges), which are the NCIs, controlling for the original IV, wheat production. They cannot reject a zero coefficient on alternative crops. Hence, they do not find an indication of a problem with the IV.

NCI tests typically require conditioning on the IV, as the NCI may be associated with the outcome even in valid IV designs. This association arises because the NCI is often associated with the IV, which in turn influences the outcome through the treatment. For example, in panels C and D of Figure 1, the NCI and the outcome are associated through the IV, even if no alternative path exists and the IV design is valid. In Nunn and Qian (2014a), orange production (the NCI) is associated with conflicts (the outcome), through their associations with wheat production (the IV).

However, if the NCI and IV are independent, conditioning on the IV is not required. In such cases, researchers can use an unconditional independence test for the null $H_0 : NC \perp\!\!\!\perp Y$, as formalized in the following theorem.

THEOREM 3: *Assume that a random variable NC satisfies the NCI assumption. If in addition $NC \perp\!\!\!\perp Z$, then if $NC \not\!\perp\!\!\!\perp Y$, either outcome independence or exclusion restriction is violated. That is, the IV design is invalid.*

Supplemental Appendix C provides a version of this theorem with control variables, in which the IV and NCI need to be conditionally independent only given the set of controls.

Situations where $NC \perp\!\!\!\perp Z$ (so, per the theorem, unconditional NCI tests may be valid) can occur when considering violations of the exclusion restriction assumption. Panel A of Figure 2 provides an example. Consider the context of Jacob, Lefgren, and Moretti (2007), who study the effect of lagged crime ($X$) on current crime ($Y$). They use lagged weather as an IV ($Z$) for lagged crime. The API variable is temporal displacement of economic activity ($U$): Lagged weather can postpone economic activity to the current period, which could in turn affect current crime ($Y$), thus violating the exclusion restriction. In this context, a different variable that displaces

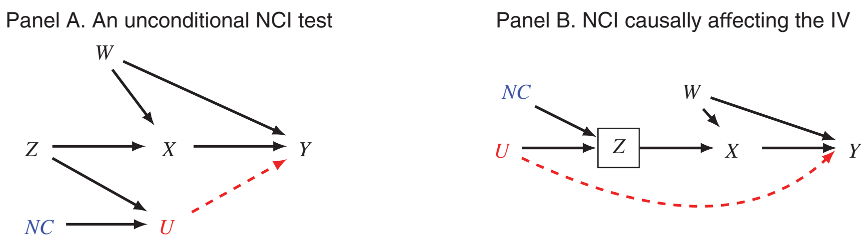


FIGURE 2. ILLUSTRATION OF SCENARIOS RELATED TO NEGATIVE CONTROL INSTRUMENT TESTS

*Notes*: Each panel represents a different scenario related to negative control instrument tests. In both scenarios, $X$ is the endogenous treatment variable, $Y$ is the outcome, $Z$ is the IV, and $W$ is a potential confounder that motivates the use of the IV. The validity of the IV design is challenged by the potential alternative paths. Panel A demonstrates an NCI scenario where conditioning on the IV is not necessary. In this example, $U$ is an unobserved API variable that poses a threat to identification. If it is related to the outcome (through the dashed arrow), the exclusion restriction is violated. An observed NCI ($NC$) that affects the API variable ($U$) can be used to evaluate the presence of the problematic association by implementing an unconditional independence test for the null hypothesis $H_0 : NC \perp\!\!\!\perp Y$. Panel B shows that a variable causally affecting the IV can serve as a valid NCI. The variable $U$ is an unobserved API variable that poses a threat to identification; if it is related to the outcome (through the dashed arrow), outcome independence is violated. The square around $Z$ symbolizes the conditioning on the IV. An observed NCI ($NC$) that affects the IV ($Z$) can be used to evaluate the presence of the suspected association by testing $NC \perp\!\!\!\perp Y|Z$. Specifically, if $NC \not\perp\!\!\!\perp Y|Z$, then the path $NC \rightarrow Z \leftarrow U \rightarrow Y$ exists, and hence, $U \not\perp\!\!\!\perp Y|Z$.

economic activity can be used as an NCI. For example, payday cycles ($NC$) are known to impact the timing of economic activity (Hastings and Washington 2010). Payday timing is independent of weather. Therefore, an association between payday and crime would imply a violation of the exclusion restriction assumption. In this case, no conditioning on $Z$ is needed.[9] By contrast, in contexts where a violation of outcome independence is suspected, the IV and the NCI are typically associated as well (as in panel C of Figure 1). Therefore, the NCI test should condition on the IV.

Nevertheless, researchers can opt to always control for the IV. Since $NC$ and $Z$ are observed, the condition $NC \perp\!\!\!\perp Z$ is empirically testable. However, researchers might opt to skip such a test and condition on the IV anyway. In a linear model, adding an additional control that is independent of the $NC$ will not affect the coefficient estimate for $NC$ asymptotically. Furthermore, if $Z$ has a causal effect on $Y$, including it in the regression can improve the precision of the estimation.

### D. *Control Variables and Functional Forms*

In many cases, the IV is believed to be valid only conditionally on certain control variables. For example, in papers that use judge assignment as an IV, the assignment of judges is quasi-random only within date and location (e.g., Kling 2006). Therefore, the independence assumption is satisfied only conditionally, and the IV design is valid only once controlling for date and location.

Formally, let $C$ be the vector of controls. Similar to the case without controls, outcome independence and exclusion restriction together imply $Z \perp\!\!\!\perp Y(x) \big| C$.

[9]The NCI assumption is that the association between payday timing and crime is driven only by economic activity timing.

Supplemental Appendix C presents the previous results when control variables are included.

When the IV is presumably valid only conditional on a vector of control variables $C$, an NCO test is a test for the null hypothesis

$$H_0 : NC \perp\!\!\!\perp Z | C. \tag{2}$$

Similarly, for NCIs, the null hypothesis is

$$H_0 : NC \perp\!\!\!\perp Y | C, Z. \tag{3}$$

While accounting for controls in an IV analysis can be done in a variety of ways (e.g., Abadie 2003), the large majority of applications use a two-stage least squares (2SLS) specification. This specification makes additional functional form assumptions. Most negative control tests used in practice adopt the same functional form as the 2SLS.

In particular, NCO tests typically adopt the functional form for how the IV depends on the control variables. To avoid excessive notation, let $C$ also denote the set of controls in a 2SLS specification.[10] Blandhol et al. (forthcoming) show that 2SLS requires the following linearity assumption.[11] This assumption is required for 2SLS to satisfy their definition of a *weakly causal estimand*.[12]

ASSUMPTION 3 (Rich Covariates): *The conditional expectation of the IV is linear in the control specification. Namely,* $\mathbb{E}[Z | C] = \tilde{\gamma}_C' C$, *for some vector* $\tilde{\gamma}_C$.

Combining the null hypothesis in (2) and rich covariates, we expect that

$$\mathbb{E}[Z | C, NC] = \tilde{\gamma}_C' C. \tag{4}$$

This equation provides a more specific null hypothesis for conditional independence testing. This hypothesis can be tested by regressing the IV on the vector of controls and the NCO. The following corollary of Theorem 1 formalizes this.

COROLLARY 1: *Assume that the random variable NC is an NCO. Let*

$$\gamma = (\gamma_C', \gamma_{NC}) = \underset{b_c, b_{NC}}{\operatorname{argmin}} \mathbb{E}[Z - b_C' C - b_{NC} NC]^2$$

*be the population-level OLS coefficient of regressing Z on C,NC. If* $\gamma_{NC} \neq 0$, *then either outcome independence*, *exclusion restriction*, *or rich covariates is violated.*

[10] The vector $C$ may include, for example, a quadratic function of one of the original controls or interactions. For ease of notation, $C$ would always include the intercept.

[11] Like Assumptions 1 and 2, the functional form assumptions refer to the assumptions of the IV design that are tested by the negative control tests rather than the assumptions needed for the validity of the negative control tests, discussed in the following corollaries.

[12] A weakly causal estimand is a positively weighted average of subgroup-specific treatment effects.

Researchers often run the reverse regression in which the NCO is the outcome variable. This practice is equivalent, as formalized in the following corollary.

COROLLARY 2: *Assume that NC is an NCO. Let*

$$\beta = (\beta_Z, \beta'_C) = \underset{b_Z, b_C}{\operatorname{argmin}} \mathbb{E}\left[NC - b_Z Z - b'_C C\right]^2$$

*be the population-level OLS coefficient of regressing NC on Z, C. If* $\beta_Z \neq 0$, *then either outcome independence, exclusion restriction, or rich covariates is violated.*

NCI tests typically adopt the functional form of the relationship between the outcome and the IV and the control variables. Specifically, NCI tests often use the same structure as the reduced-form equation. Therefore, they implicitly make the following assumption.

ASSUMPTION 4 (Correctly Specified Reduced Form (CSRF)): *The conditional expectation of the outcome is linear in the IV and the control variables. Namely,* $\mathbb{E}[Y|Z,C] = \tilde{\theta}_Z Z + \tilde{\theta}'_C C$, *for some* $\tilde{\theta}_Z$ *and vector* $\tilde{\theta}_C$.

Combining the null hypothesis (3) with the CSRF assumption, we expect that

$$\mathbb{E}[Y|Z,C,NC] = \tilde{\theta}_Z Z + \tilde{\theta}'_C C. \tag{5}$$

This equation also provides a more specific null hypothesis, which can be tested with OLS. The following corollary of Theorem 2 shows that such an OLS jointly tests IV violations due to an alternative path and CSRF.

COROLLARY 3: *Assume that the random variable NC is an NCI. Let*

$$\theta = (\theta_Z, \theta'_C, \theta_{NC}) = \underset{b_Z, b_c, b_{NC}}{\operatorname{argmin}} \mathbb{E}\left[Y - b_Z Z - b'_C C - b_{NC} NC\right]^2$$

*be the population-level OLS coefficient of regressing Y on Z, C, NC. If* $\theta_{NC} \neq 0$, *then either outcome independence, exclusion restriction, or CSRF is violated.*

Corollaries 1, 2, and 3 imply that, in the tests discussed, the null hypothesis can be rejected in IV designs that satisfy outcome independence and exclusion if functional form assumptions are violated. For NCO tests, the null can be rejected because the rich covariates assumption is not satisfied. In such cases, researchers can still estimate a causal effect by modifying the functional form or using methods other than 2SLS (Blandhol et al. forthcoming). For NCI tests, rejection may stem from a violation of the CSRF assumption. However, unlike rich covariates, CSRF is not a necessary assumption for 2SLS analysis. Hence, negative control tests sensitive to this assumption could reject perfectly valid IV designs.

For example, an NCI test can reject the null in designs where the IV is randomly assigned due to CSRF violation. Random assignment guarantees that outcome independence and rich covariates hold. Assuming the exclusion restriction holds, the design is valid. However, CSRF could still be violated if the IV has a nonlinear effect on the outcome or a heterogeneous effect across control vector values. In such cases, an NCI test may reject the null in (5), despite the design being valid.

## III. Implementation Guidance

This section offers guidelines for the implementation of negative control tests. We recommend that researchers follow four steps, summarized in Appendix Figure 1. First, when possible, researchers should articulate specific threats to the validity of the IV design and characterize the alternative paths variables, as discussed in Section IIA. Second, researchers should survey available data to identify suitable negative controls—variables that can serve as proxies for the unobserved alternative path variables. These proxies should satisfy the NCO or NCI assumption (see Definitions 3 and 4). Examples are discussed in Section IIIB.

Third, researchers should choose a statistical test for independence between the NCO and the IV or between the NCI and the outcome, conditioning on the IV. For IV designs that require conditioning on a set of controls, negative control tests should also condition on these controls. Section IIIC discusses particular test specifications, their validity, and the assumptions they test, which in some cases also include functional form assumptions. Fourth and finally, researchers should interpret the result and conduct further diagnostics if the test rejects the null, as discussed in Section IIID.

To illustrate these recommendations, we apply them to IV designs used in prior work. We chose four widely cited papers published in the *American Economic Review* with publicly posted replication data. We use Autor, Dorn, and Hanson (2013a) and Deming (2014a) to discuss NCO tests and Ashraf and Galor (2013a) and Nunn and Qian (2014a) to discuss NCI tests.[13] Table 3 summarizes the IV designs in these papers and the negative controls they used in their falsification tests. Supplemental Appendix E provides additional details on our analyses.

### A. *Potential Threats*

Two guiding questions can help researchers characterize potential violations of IV validity. This characterization can assist in selecting appropriate negative control variables and determining which hypothesis should be tested. The first question is whether the primary concern is a violation of outcome independence (Assumption 1) or the exclusion restriction (Assumption 2). Both types of violations introduce an alternative path between the IV and the outcome.

Outcome independence is violated when the alternative path is through some factor that affects both the IV and the outcome. As previously discussed, Martin and Yurukoglu (2017) examine whether Fox News viewership influences Republican vote shares using cable channel positions as an IV. The concern is that cable companies may place Fox News in lower channel numbers in conservative locations,

[13] Ashraf and Galor (2013a) and Nunn and Qian (2014a) are the two most cited AER papers published since 2013 that use an NCI test. Autor, Dorn, and Hanson (2013a) is the most cited AER paper published after 2013 that uses an NCO test. Deming (2014a) was selected to demonstrate how our proposed follow-up analysis can be used to diagnose and correct problems with the IV design in Section IIID. In all cases, we use the original replication data: Deming (2014b); Autor, Dorn, and Hanson (2013b); Ashraf and Galor (2013b); and Nunn and Qian (2014b).

Table 3—Papers Used to Illustrate Applications of Negative Control Tests

| | Original variable description | | | | # of negative controls | |
|---|---|---|---|---|---|---|
| | Outcome (1) | Treatment (2) | IV (3) | Negative controls (4) | Original paper (5) | Our analyses (6) |
| *A. Negative control outcome tests* | | | | | | |
| Autor, Dorn, and Hanson (2013a,b) | Local labor market outcomes | Shift-share using import shares from China | Shift-share using other countries' import shares from China | Lagged local labor market manufacturing employment | 3 | 52 |
| Deming (2014a,b) | Student test scores | Value added of the school | School assignment lottery, interacted with school VA | No negative control analysis in the original paper | – | 37 |
| *B. Negative control instrument tests* | | | | | | |
| Ashraf and Galor (2013a,b) | Level of economic development | Population genetic diversity | Migratory distance from East Africa | Migratory distance from alternative locations not associated with genetic diversity (e.g., London) | 3 | 3 |
| Nunn and Qian (2014a,b) | Level of civil conflict | Receipt of US food aid | Variation in US wheat production | Variation in US production of crops not associated with aid (e.g., oranges) | 10 | 10 |

*Notes:* This table provides contextual details on the papers we use for our negative control application examples in Table 4 and Table 5. Columns 1–3 specify which outcome, treatment, and IV were used in these papers, respectively. Column 4 depicts the negative controls used in the original analysis, and column 5 presents the number of negative controls used. Column 6 presents the number of negative controls used in the analysis for Table 4 and Table 5, including all the original negative controls, as well as additional valid negative controls we found in the original data.

where voters lean Republican regardless. Violations of outcome independence are illustrated in panels A and C of Figure 1.

The exclusion restriction is violated when the IV affects the outcome through channels other than its effect through the treatment. For example, as previously discussed, in Angrist and Evans (1998), the sex composition of the first two children (the IV) may affect female labor force participation (the outcome) not only through its effect on family size (the treatment) but also through its effects on household expenditures due to hand-me-downs. Exclusion restriction violations can occur even with randomly assigned IVs, as in a randomized controlled trial. Violations of the exclusion restriction are illustrated in panels B and D of Figure 1.

The second question is whether the threat (the alternative path) operates through an APO or an API variable. That is, does the alternative path variable have a known association with the outcome, and the concern is that it may also be associated with the IV? Or does it have a known association with the IV, and the concern is that it may also be related to the outcome? In the first case, the alternative path operates through an APO variable; in the second, it operates through an API variable. In the context of Martin and Yurukoglu (2017), unobserved conservativeness is an APO variable; it certainly influences voting for Republican candidates (the outcome), yet it is unclear whether it is also associated with Fox News channel placement (the

IV). By contrast, in Nunn and Qian (2014a), weather conditions are an API variable; they surely affect wheat production (the IV), and the concern is that they may also directly affect the conflicts in aid recipient countries (the outcome). Note that both outcome independence and exclusion restriction can be violated through either APO or API variables. Panels A and B of Figure 1 demonstrate this for APO variables, and panels C and D for API variables.

These two questions can assist researchers in finding relevant negative control variables and choosing the right negative control tests. For APO variables, researchers should look for NCOs and test their association with the IV. For API variables, researchers should look for NCIs and test their association with the outcome, conditionally on the IV. The type of violation (outcome independence or exclusion restriction) can guide the search for relevant negative control types.

### B. *Choosing Negative Controls*

In this section, we discuss different types of negative controls, both commonly used in practice and novel ones suggested by the results presented earlier.

#### *Common Types of Negative Control Outcomes*

**Predetermined Variables:** Variables fixed before the IV is determined are frequently used in NCO tests. Common examples include lagged outcome variables and demographic characteristics such as gender, race, and age. Predetermined variables are useful for testing outcome independence. In some cases, researchers may choose to use predetermined variables as NCOs even without clear knowledge of which exact APO variable they proxy for. If the IV is associated with a predetermined variable, it could imply that it is affected by something that affects the outcome as well.

However, not every predetermined variable is a valid NCO. First, NCOs need to satisfy U-comparability (Definition 3); predetermined variables that are completely unrelated to the outcome are uninformative and should not be used. Second, not all predetermined variables satisfy the NCO assumption. In particular, certain predetermined variables may influence the IV, even if the underlying IV design is valid. For example, when the IV is the child's quarter of birth (as in Angrist and Krueger 1991), the parents' quarter of marriage is not a valid NCO, as it likely influences the child's quarter of birth, even if the design is valid.

When IVs are assumed to be quasi-randomly assigned (e.g., lotteries), they cannot be affected by predetermined variables. Researchers could therefore use predetermined variables as NCOs to evaluate the claim that the IV is quasi-random.[14] If the IV is associated with a predetermined variable, it is unlikely to be quasi-random, and hence, outcome independence might not hold. For example, we found multiple predetermined variables in the replication data from Deming (2014b), which uses an IV constructed based on school lotteries. We use these predetermined variables as NCOs to evaluate outcome independence. We have found an association of the

[14] In this case, predetermined variables are NCOs based on the more general Definition A9, which allows for the NCO to be directly associated with the IV (only) if the IV is not quasi-random as claimed.

IV with the predetermined variables. In particular, we found that the construction of the IV involved nonrandom components that require additional controls; see Section IIID.

Predetermined variables are also useful NCOs when the IV is not quasi-random. For example, Autor, Dorn, and Hanson (2013a) use a shift-share IV for commute-zone exposure to Chinese imports to evaluate their impact on employment. To evaluate this IV, they use predetermined local labor market manufacturing employment as NCOs. The concern is that since industry exposure to Chinese imports is nonrandom, it might be associated with other labor market conditions, which in turn could be associated with the outcome (e.g., Chinese imports were more pronounced in regions with industries that were declining in Western countries regardless). Such an association would violate outcome independence. We found many additional predetermined variables in the original paper's replication data that could proxy for latent local labor market conditions (e.g., past unemployment). These variables can also serve as NCOs. The NCO assumption requires that any association of the IV with the predetermined variables used as NCOs is driven by an APO variable, that is, by something that also affects the outcome. This would be violated if, for example, Chinese import penetrated industries due to economic factors that were only relevant in the past and are no longer relevant in the studied period.

**IV Leads and Lags:** Certain IVs are predicated on serendipitous or chance occurrences ("strokes of luck"). Because the occurrence of such unexpected shocks should be independent across time, leads and lags of the variable used as the IV can serve as NCOs. For example, Jäger, Heining, and Lazarus (2024) use a worker's premature death as an IV for employee turnover, under the assumption that such deaths occur randomly across firms. A potential concern is that deaths are nonrandom and reflect riskier conditions in the firm (the APO variable) that directly impact wages (the outcome). This would violate outcome independence. To rule this out, Jäger, Heining, and Lazarus (2024) use subsequent premature deaths in the same firm as an NCO that proxies for potentially unobservable riskier conditions. The NCO assumption here stipulates that given the risk conditions, premature deaths in different time periods should be conditionally independent. A recurring pattern of premature deaths would cast doubt on the assumption that such deaths occur randomly across firms.

**Alternative Outcomes:** Alternative or unrelated outcomes can also serve as NCOs for two different types of APO variables. First, APO variables can potentially affect the IV, forming an alternative path that violates outcome independence (as in panel A of Figure 1). Alternative outcomes that are affected by the same APO variable can then serve as NCOs. For example, Chetty, Friedman, and Rockoff (2014) leverage teachers' moves between schools to evaluate middle school teacher value added measures. The concern is that high-quality teachers may tend to move to schools that experience simultaneous improvements in student quality. Here, the APO variable is the unobserved changes in school quality. To evaluate this threat, Chetty, Friedman, and Rockoff (2014) use as NCOs test scores from subjects not taught by the teacher in question. If the NCO test finds that teacher quality is associated with better outcomes in subjects they do not teach, it would cast doubt on the design's validity. Chetty, Friedman, and Rockoff (2014) focus on middle school teachers as opposed to elementary school teachers who teach multiple topics. This is because the NCO assumption requires that the IV will not affect the alternative outcome directly. The IV should also not affect alternative outcomes indirectly via the treatment or outcome.

The second type of APO variables potentially violates the exclusion restriction. The concern is that the IV affects an additional factor (the APO variable), which in turn affects the outcome (as in panel B of Figure 1). In the example of Angrist and Evans (1998), the concern is that same-sex sibship IV might affect female labor supply due to hand-me-downs, thus forming an alternative path. To evaluate this concern, Rosenzweig and Wolpin (2000) check the association of same-sex sibship with an alternative outcome: clothing expenditure.[15]

### *Common Types of Negative Control Instruments*

**Variables Similar to the IV That Do Not Affect the Treatment:** Researchers often choose NCIs that are similar to the IV but are presumed not to influence the treatment variable. These NCIs typically test outcome independence. They usually share many similarities with the IV and are thus likely to be associated with the API variable. For example, as previously discussed (and illustrated in panel C of Figure 1), Nunn and Qian (2014a) use US wheat production as their IV for US aid and consider the US production of other crops unrelated to US aid (e.g., oranges) as NCIs. These variables are similar, as they are both affected by the same API variables, such as weather conditions. A similar type of NCI exists in Ashraf and Galor (2013a), who study the impact of genetic diversity on economic development. They use distance from Addis Ababa as their IV for genetic diversity. To test the IV design, they construct NCIs from distance from other cities unrelated to the human origin hypothesis (e.g., London, Tokyo).

In some cases, researchers generate variables similar to the IV on their own. They construct a variable in a similar way to how the IV was constructed but remove the impact on the treatment. For example, De Giorgi, Frederiksen, and Pistaferri (2020) study the effect of peers' consumption on own consumption. As an IV, they use economic shocks to firms of distant peers. This IV will be associated with shocks to large firms (as statistically, they are more likely to affect all workers, including distant peers), which could potentially affect the outcome in other ways. To test this, they use an NCI which they call a "placebo" IV; they calculate the same IV when replacing the real allocation of workers to employers with a random allocation, keeping firm sizes constant.

**IV Leads:** Future instances of the IV (IV leads) can often serve as effective NCIs for testing outcome independence. IV leads are often affected by the same variables as the IV, and specifically by an API variable, as in panel C of Figure 1.

[15] The NCO is the APO variable itself, so the NCO assumption is trivially satisfied (see Section IIB).

For example, Moretti (2021) studies the effect of the size of regional high-tech clusters on local inventors' productivity, measured via their number of new patents. As an IV, he uses predicted cluster size based on the expansion of local firms outside the cluster. One possible identification threat is the sorting of productive inventors into firms that are expected to grow. The API variable is firms' anticipated growth, which affects the IV. The concern is that this same anticipation could also attract productive innovators from other regions. This will increase the outcome due to innovators' sorting and not because of the effect of the cluster size on the local innovators. Anticipated growth potentially affects more than one period; therefore, this API variable is likely to affect future predicted cluster sizes (i.e., IV leads) as well. Moretti then ascertains that future predicted cluster size is (conditionally) independent of productivity.

To satisfy the NCI assumption, the outcome must not influence future realizations of the IV. In the example of Moretti (2021), this implies that regional productivity cannot affect the expansion of local firms in other locations.

When practitioners observe IV leads, they need to consider whether they expect these variables to be associated with the IV. When the IVs are expected to be autocorrelated (as in Moretti 2021), an NCI strategy can be used. When the IVs are presumably independent, an NCO strategy can be applied, as discussed in the previous section.

*Underutilized Types of Negative Control Instruments*

**Causes of the IV:** In some cases, researchers may suspect violations of outcome independence through API variables that affect the IV and potentially also affect the outcome. In these cases, researchers can use variables that causally affect the IV as NCIs. In the example from Angrist and Krueger (1991), the parents' quarter of marriage influences the child's quarter of birth (the IV) and qualifies as a valid NCI. In this case, API variables are factors that influence a child's quarter of birth and are suspected to affect wages (the outcome). For example, certain parental occupations exhibit seasonal patterns that affect the timing of births and may independently influence their children's later earnings.

Panel B of Figure 2 illustrates how such NCIs work. When both the API variable and the NCI influence the IV, they are associated conditional on the IV (in this DAG, the IV acts as a *collider*; Pearl 2009b). If, conditional on the IV, the NCI is also associated with the outcome, it implies that a path exists between the NCI and the outcome through the IV and the API variable. Therefore, the API variable affects the outcome, thus violating outcome independence.

To satisfy the NCI assumption, these NCIs should have no association with the outcome other than through the IV.[16] In particular, the NCI cannot directly affect either the treatment or the outcome. In the example of Angrist and Krueger (1991), using parents' quarter of marriage as an NCI requires assuming that marriage timing does not directly influence child schooling or wages.

[16] If such an association does exist, these variables must be used as controls.

**IV Side Effect Proxies:** An IV that not only affects the treatment but also produces a side effect may violate the exclusion restriction. This occurs if the side effect also affects the outcome. In such cases, proxies for the side effect (the API variable) may serve as NCIs.

Panel D of Figure 1 illustrates a scenario where the IV influences the NCI through the API variable (therefore, the NCI is itself a side effect). For example, in the previously discussed context of Jacob, Lefgren, and Moretti (2007), lagged weather ($Z$) serves as an IV for lagged crime ($X$) to study its impact on current crime ($Y$). Lagged weather also creates intertemporal displacement of economic activity (the API variable $U_4$). The concern is that intertemporal displacement of economic activity affects subsequent crime, thus violating the exclusion restriction. To evaluate whether this alternative path exists, Jacob, Lefgren, and Moretti (2007) use traffic patterns, a proxy for economic activity, as an NCI ($NC_4$). Testing whether traffic patterns are independent of crime, conditional on lagged weather, constitutes an NCI test for this alternative path. If an association is found, it suggests that the exclusion restriction is violated, as weather influences crime not only through past crime but also through displaced economic activity. Alternatively, panel A of Figure 2 presents another type of side effect proxy that influences the API variable rather than being influenced by it. For Jacob, Lefgren, and Moretti (2007), that could be other factors that displace economic activity (e.g., payday schedule; see the discussion of this example in Section IIC).

*Power Considerations When Choosing Negative Controls*.—Negative control variables must satisfy U-comparability. This implies that these variables are indeed associated with the alternative path variable. Some negative controls might satisfy this condition but have only a weak association with the alternative path variable. In this case, if the IV design is not valid, the association between the NCO and the IV or the NCI and the outcome would be difficult to detect without having a large dataset. Therefore, power considerations suggest excluding negative control variables that have only a weak association with the alternative path variable, as they can lower test power. This mirrors the effect of irrelevant control variables in OLS. This issue is especially acute for NCI variables that are intentionally similar to the original IV. Such variables are often strongly associated with the IV but only weakly associated with the API variable, conditional on the IV.

### C. *Choosing a Statistical Test*

As discussed in Section IIC, negative control tests assess whether a negative control variable is (conditionally) independent of the IV or outcome. The choice of statistical test for conditional independence should take into account three primary considerations: the estimation method (e.g., 2SLS), the anticipated functional form of the relationship between the negative control and either the IV or the outcome (and the controls), and statistical power. This section discusses both commonly used and underused conditional independence tests and presents examples using replication data from existing work.

*Commonly Used Tests*.—The most commonly used NCO falsification tests are based on the original IV reduced-form equation, $Y = \alpha_Z Z + \alpha_C' C + \epsilon$, but replace the outcome with an NCO (e.g., past outcomes). That is, the test estimates the model

$$NC = \beta_Z Z + \beta_C' C + \epsilon_{NC} \tag{6}$$

and evaluates the null hypothesis $H_0 : \beta_Z = 0$. This test evaluates outcome independence or exclusion restriction, as well as the rich covariates assumption (Corollary 2). Because in 2SLS the rich covariates assumption is necessary for causal interpretation, this test provides useful information regarding both the validity of the IV design and of the 2SLS specification. Since this test uses the same inferential framework as the original study, it can also expose errors in the inference method (Eggers, Tuñón, and Dafoe 2024).

When multiple negative controls are available, model (6) can be estimated separately for each negative control. The null hypothesis is that the coefficient $\beta_Z$ equals zero in *all* the regression equations. To test this null hypothesis with a single test statistic, the researcher should account for the covariance between the estimated $\beta_Z$ coefficients across the different model equations. This could be done, for example, by employing a generalized method of moments that combines all estimating equations together.

An alternative approach is to jointly incorporate multiple negative controls using the model

$$Z = \gamma_{NC}' NC + \gamma' C + \epsilon_Z, \tag{7}$$

where $NC$ now represents a vector of NCOs.[17] An $F$-test can be used to evaluate the null hypothesis $H_0 : \gamma_{NC}' = 0'$ under standard assumptions.[18]

For NCI tests, a similar approach applies, but the outcome is regressed on the NCI, controlling for the IV. Specifically, the NCI (denoted again NC) is added to the reduced-form equation

$$Y = \theta_Z Z + \theta_C' C + \theta_{NC} NC + \epsilon_Y, \tag{8}$$

and the null hypothesis is $H_0 : \theta_{NC} = 0$. A common mistake in practice is omitting the IV from this regression (see Section I). Doing so can lead researchers to reject the null, even when the IV design is valid. The IV can be omitted in the (rare) event that the NCI is independent of the IV.

This NCI test can still reject the null for valid IV designs (even when conditioning on the IV). Corollary 3 shows that even if the IV design is valid, the null could still be rejected due to a violation of CSRF (Assumption 4). As discussed in Section IID, the CSRF assumption is not necessary for causal identification and can be violated even under random assignment.[19] Therefore, tests based on Model (8) may reject

[17] In most applications, if a vector of negative controls is associated with the IV, at least one of its components will be as well. Supplemental Appendix D.5 provides a theoretical counterexample, but such cases are unlikely in practice, as small parameter changes would reverse the result.

[18] With robust standard errors, the common implementation calculates the Wald statistic, divides it by the degrees of freedom, and calculates a $p$-value from an $F$-distribution.

[19] An exception is binary IV without controls, in which case CSRF is always satisfied.

the null hypothesis, even though the IV design can identify a causal estimand with 2SLS. To relax this sensitivity to linearity assumptions, researchers may opt for semiparametric or nonparametric tests, which are discussed next.

*Additional Tests*.—The tests discussed above inherit the basic functional form assumption as in the 2SLS specification. The NCO tests replace the outcome with the NCO in the reduced-form equation or posit the reverse linear model for the IV. The NCI tests include the NCI additively in the reduced form.

Researchers should consider replacing the functional form assumptions in two cases. First, tests with more flexible functional forms are useful when the functional form assumptions are undesirable. In particular, when using an NCI, researchers should test for nonlinear associations whenever possible to avoid testing the CSRF assumption, which is unnecessary.[20]

Second, more flexible tests are useful when researchers suspect a nonlinear association between the NCO and the IV or between the NCI and the outcome. Such nonlinear associations also indicate a violation of the IV assumptions and should therefore be tested whenever possible. For example, in studies using crops as an IV (as in Nunn and Qian 2014a), one might be concerned that extreme weather conditions, such as unusually high or low temperatures, could affect both crop yield and conflict incidence (the outcome). In this case, Model (8) might fail to detect a nonmonotonic relationship between an average temperature NCI and the outcome.

To estimate more complex functional forms, researchers can include higher-order polynomial terms or interactions in regression models. Alternatively, they can use semi- or nonparametric tests. The next section discusses a few examples, which we implement in our application examples.

When using estimation methods other than 2SLS, researchers should consider more general conditional independence tests that assess relationships beyond mean independence. To this end, a variety of other conditional independence tests can be considered (see Heinze-Deml, Peters, and Meinshausen 2018 and Li and Fan 2020 for recent surveys). The choice of test depends on the specific context, as there is no uniformly optimal conditional independence test.[21] Naturally, more complex tests require larger datasets or low-dimensional covariates for reliable implementation. Therefore, these tests may be less informative in small samples or in settings with many covariates.

*Examples of Applications*.—Table 4 presents the NCO tests results using data from Autor, Dorn, and Hanson (2013b) and Deming (2014b). Column 2 shows that a single NCO test fails to reject the null hypothesis in both cases.[22] However, a joint $F$-test in column 3 demonstrates that using multiple NCOs leads to a rejection of

[20] By contrast, the previously discussed linear NCO tests examine the rich covariates assumption, which is necessary for 2SLS. Therefore, tests based on Models (6) or (7) are often preferable.

[21] Shah and Peters (2020) show that for continuous distributions, there is no conditional independence test that is uniformly valid and is simultaneously powerful against any conditional dependence types.

[22] For Autor, Dorn, and Hanson (2013a), column 1 replicates the original NCO test from their paper, without control variables. They find the IV is significantly associated with the lagged outcome, albeit with the opposite sign from the main analysis. This association becomes insignificant when all controls are included.

TABLE 4—ILLUSTRATIVE APPLICATIONS OF NEGATIVE CONTROL OUTCOME TESTS

| | Original analysis | Alternative analyses | | | | |
|---|---|---|---|---|---|---|
| Type of test: | Single NCO linear model without controls (1) | Single NCO linear model (2) | $F$-test (3) | GAM, linear controls (4) | GAM, nonlinear controls (5) | Number of observations (6) |
| Autor, Dorn, and Hanson (2013a,b) | 0.004 | 0.593 | $<$ 0.001 | 1.000 | 1.000 | 722 |
| Deming (2014a,b) | – | 0.686 | $<$ 0.001 | $<$ 0.001 | $<$ 0.001 | 2,343 |

*Notes:* This table presents $p$-values from different NCO tests using data from Autor, Dorn, and Hanson (2013a,b) and Deming (2014a,b). Column 1 replicates one of the original falsification analyses, in which Autor, Dorn, and Hanson (2013a,b) replaced the outcome with the NCO in the same 2SLS specification as their main analysis (Autor, Dorn, and Hanson (2013a,b), Table 2, Part II). Deming (2014a,b) conducted no falsification tests. Columns 2–5 report $p$-values obtained from additional tests that include the same controls as in the most exhaustive specification of the original analyses. Column 2 reports a test using one NCO, where the outcome is replaced with the NCO in the reduced-form regression. For Autor, Dorn, and Hanson (2013a,b), the single NCO is the lagged outcome (in 1970), which is the same NCO reported in column 1; for Deming (2014a,b), this is lagged test scores (2002). Column 3 uses an $F$-test (Model (7)) with all NCOs jointly. Columns 4 and 5 use GAM tests with linear and smoothed controls, respectively (Model (9)).

the null. This result illustrates how the inclusion of additional NCOs may enhance test power.

To examine nonlinear associations, we implement the common semiparametric approach of Generalized Additive Models (GAMs; see Hastie and Tibshirani 1990; Wood 2006). We express the IV as an additive combination of smooth functions of the controls ($C$) and NCOs ($NC$):

$$Z = \sum_j f_j(C_j) + \sum_k g_k(NC_k) + \epsilon_Z, \tag{9}$$

where $f_j, g_k$ are smooth functions estimated via splines. We test whether $g_k = 0$ for all $k$ to assess whether the NCOs are conditionally independent of the IV.[23]

To test rich covariates (Assumption 3) within this framework, researchers can restrict $f_j$ to be linear (i.e., $f_j(C_j) = \gamma_j C_j$) while allowing $g_k$ to remain nonlinear. Such a model is useful when using 2SLS, which requires rich covariates, while suspecting a strong nonlinear association between the NCO and the IV. Columns 4 and 5 in Table 4 implement a GAM test with and without assuming linearity in the controls. The GAM test rejects the null in Deming (2014a) but not in Autor, Dorn, and Hanson (2013a). One possible explanation is that the sample size in Autor, Dorn, and Hanson (2013a) is too small to estimate the richer nonlinear GAM model.

Table 5 presents the NCI test results for Nunn and Qian (2014a) and Ashraf and Galor (2013a). Columns 1 and 2 show the results of regressing the outcome on the NCI, both with and without conditioning on the IV. In both studies, conditioning on the IV is necessary. For Nunn and Qian (2014a), failure to condition leads to rejection of the null, suggesting a false rejection due to the path between the NCI and the outcome through the IV. For Ashraf and Galor (2013a), the null is not rejected in

[23] A similar GAM test can be implemented for NCI tests, by adding smooth functions to Model (8).

Table 5—Illustrative Applications of Negative Control Instrument Tests

| | Without conditioning on the IV | With conditioning on the IV | | |
|---|---|---|---|---|
| | Single NCI linear model (1) | Single NCI linear model (2) | $F$-test (3) | Number of observations (4) |
| Nunn and Qian (2014a,b) | 0.007 | 0.123 | 0.138 | 4,572 |
| Ashraf and Galor (2013a,b) | 0.234 | 0.778 | 0.994 | 145 |

*Notes:* This table presents $p$-values from different NCI tests using data from Nunn and Qian (2014a,b) and Ashraf and Galor (2013a.b), applying their original sets of NCIs (10 and 13 NCIs, respectively). Column 1 shows a linear NCI test with a single NCI that, inappropriately, does not condition on the IV. The NCI with the lowest $p$-value is shown (grape production for Nunn and Qian (2014a,b) and distance from Mexico City for Ashraf and Galor (2013a,b)). Columns 2 and 3 condition on the IV: Column 2 implements a proper linear NCI test (Model 8) using the same NCI as column 1; column 3 presents an $F$-test for all NCIs jointly.

either case, possibly due to limited statistical power. Column 3 of Table 5 presents the results from joint $F$-tests using all NCIs. In this case, the null is not rejected.[24]

### D. *Interpreting the Test Results*

*Rejection of the Null.*—Rejecting the null hypothesis in a parametric linear negative control test, such as an $F$-test, may indicate a violation of the IV assumptions or failure of the linearity assumption (rich covariates for NCO, CSRF for NCI). Researchers can further investigate by directly testing the linearity assumption (e.g., using a Regression equation specification Error Test (RESET) with control variables only; Ramsey 1969). With sufficient sample size, semi- or nonparametric tests can also test outcome independence or the exclusion restriction without relying on strict functional form assumptions.

When using multiple negative controls, identifying which ones drive the rejection can provide important insights. A useful diagnostic tool that can highlight potential alternative paths is a scatterplot showing each negative control's correlation with the IV ($y$-axis) against its correlation with the outcome ($x$-axis). When control variables are included in the IV specification, these correlations should be calculated after residualization.

Figure 3 displays such diagnostics for Deming (2014a). Deming uses predicted school value added, based on school lotteries, to evaluate school value added measures. Specifically, the IV is the value added of the student's preferred school if they won the lottery and of their neighborhood school if they did not.[25] The treatment is the value added of the school the student attended, and the outcomes are test scores in math and reading. Control variables include lagged test scores from the

[24] Due to the small sample sizes relative to the number of control variables, proper estimation of GAM models is infeasible for both studies. For Nunn and Qian (2014a), we estimate a GAM model assuming linear controls; see Supplemental Appendix E.3.

[25] Based on the replication code, the IV is defined as $IV_i = L_i VAM_i^1 + (1 - L_i) VAM_i^N$, where $L$ is the binary school lottery outcome, $VAM^1$ is the value added of the first-choice school, and $VAM^N$ is the value added of the default neighborhood school.

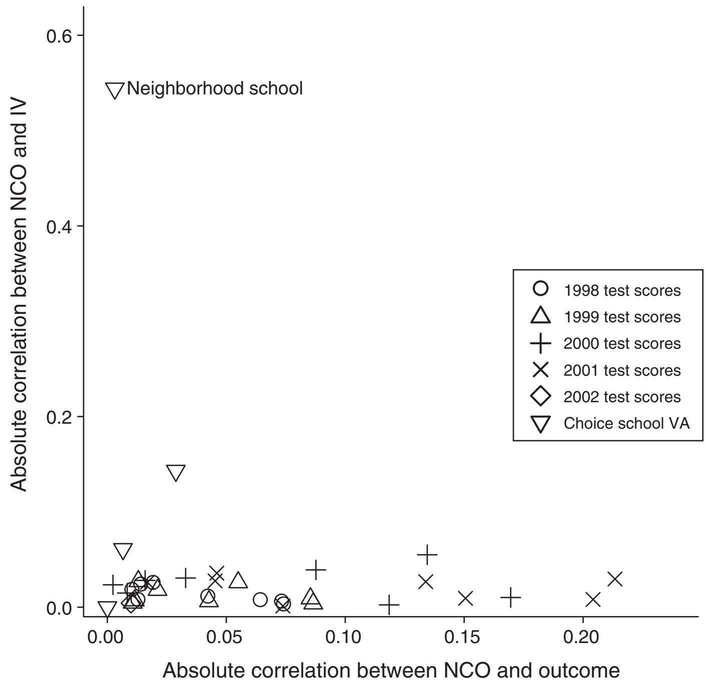


FIGURE 3. CORRELATIONS OF NCOs WITH THE IV AND THE OUTCOME IN DEMING (2014a,b)

*Notes:* This figure displays a scatterplot of the absolute value of the residualized correlation of different NCOs with the IV (on the *y*-axis) and the outcome (on the *x*-axis) in Deming (2014a,b). The NCOs, the IV, and the outcome were first residualized by regressing them on all control variables and lottery fixed effects. Each observation is one NCO. For presentation purposes, NCOs are grouped into categories denoted by marker shape. The different year markers refer to groups of students' test scores from that year. The VA marker denotes the value added of the schools listed by students as their first, second, and third submitted preferences, as well as their neighborhood school's VA (labeled Neighborhood school). See Supplemental Appendix E.2 for details.

year 2001–2002 as well as lottery fixed effects (i.e., a categorical variable for every choice of school ranking).

This IV satisfies outcome independence only when controlling for the relevant school value added measures and the probability of winning the lotteries. The scatterplot reveals that the NCO with the strongest (residualized) correlation with the IV is the value added of the neighborhood school. This occurs because the original 2SLS analysis does not control for the neighborhood school value added. This diagnostic thus identifies a fixable problem in the IV construction. This problem is resolved when using the original lottery results as an IV, as outcome independence is then satisfied.[26] This analysis serves merely as an illustration. More broadly, tests following specification changes are potentially subject to criticism, as they may not adequately control the type I error.

One caveat of this diagnostic exercise is that correlations with negative controls may not directly reflect the strength of alternative paths. Negative controls are proxies, and so, generally, the strength of their association with the IV or outcome

[26] While estimates using the original lottery are noisier, they do not alter the main conclusions.

depends on the strength of their association with the alternative path variables. Thus, weak associations between a negative control and the IV or outcome might still mask strong alternative paths. Correlations may also fail to capture nonlinear relationships between negative controls and the IV or outcome.

*Nonrejection of the Null*.—As discussed in Section II, failure to reject the null does not imply that the IV is valid. Two concerns remain. First, the IV may still be invalid due to alternative path variables not captured by the NCO or NCI used in the test. For example, a quasi-random allocation to teachers that is independent of students' neighborhoods could still be associated with students' abilities within neighborhoods. Second, an invalid IV design may pass the test due to limited statistical power.

## IV. Conclusion

This paper provides a thorough examination of the assumptions underlying negative control tests for IV designs. Our analysis clarifies existing practices and emphasizes several issues of direct practical relevance. First, most current implementations of NCI tests fail to condition on the original IV, which could lead to the unwarranted rejection of valid IV designs. Second, common negative control tests assess not only the outcome independence and exclusion restriction assumptions but also assess specific functional form assumptions. Because these assumptions are replaceable and sometimes unnecessary, researchers should distinguish between the essential IV identification conditions and the ancillary functional form assumptions when interpreting and considering additional tests. Third, our analysis clarifies what variables can serve as negative controls. These include variables that are rarely used in practice, such as variables that causally affect the IV. Moreover, in some cases, negative control variables are readily available in researchers' datasets and should be used to construct more powerful negative control tests. We hope this paper will foster a more systematic and efficient use of negative control falsification tests in empirical IV designs.

While this paper focused on the role of negative controls in testing the IV assumptions, negative controls can also be used for other purposes. As discussed, when the IV design is valid, NCOs can be used to test functional form assumptions. NCOs can also improve the estimation precision, for example, when included as control variables. Since NCOs are associated with the outcome but not with the IV, including them as controls can improve precision. By contrast, NCIs cannot be used similarly. They do not test a necessary functional form assumption, as discussed in Section IID. Moreover, including NCIs in the reduced form is likely to decrease precision because they are associated with the IV and not with the outcome.[27]

[27] In case of heteroskedasticity, at face value, NCIs may be used to improve precision in valid designs by including additional moment conditions for their orthogonality with the error as in Cragg (1983). However, this can also be achieved by including other functions of the IV.

## Appendix

1) **Articulate Potential Threats to IV Design**: Characterize potential threats as *alternative paths* between the instrument $(Z)$ and the outcome $(Y)$ through an *alternative path variable* $(U)$:
    a) Which assumption is potentially violated, outcome independence (Assumption 1) or exclusion restriction (Assumption 2)?
    b) What is the type of the alternative path variable $(U)$?
        - Alternative Path Outcome (APO): $U$ is associated with the outcome $(Y)$, and the researchers are concerned about a potential association with the IV $(Z)$ (Definition 1).
        - Alternative Path Instrument (API): $U$ is associated with the IV $(Z)$, and the researchers are concerned about a potential association with the outcome $(Y)$ (Definition 2).

   $*$Both outcome independence and exclusion restrictions can be violated through either APO or API variables. See Section IIA for definitions and Figure 1 for illustrations.
2) **Survey Available Data to Identify *Negative Control* Variables**:
    - Negative Control Outcomes (NCOs) proxy for APO variables.
    - Negative Control Instruments (NCIs) proxy for API variables.

   Section IIB discusses the assumptions these proxies need to satisfy. Examples are discussed in Section IIIB.
3) **Select Negative Control Test specification**:
    - NCO tests: Assess the independence between the IV and NCO variables. Regress IV on NCOs and controls to also test Rich covariates (Assumption 3).
    - NCI tests: Evaluate the independence between the NCI variables and the outcome, typically conditional on the IV.

   For IVs valid conditional on control variables, negative control tests should also condition on these controls.
4) **Interpret Results**:
   If the null is rejected:
    - Investigate violations of IV design or functional form assumptions separately.
    - For multiple NCs, identify which are most predictive of the IV or outcome to diagnose specific threats.

   If the null is not rejected:
    - Consider insufficient power or weak negative controls.
    - Note that untested alternative paths may still exist.

Appendix Figure 1. Steps for Implementing Negative Control Tests

## REFERENCES


**Abadie, Alberto.** 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113 (2): 231–63.

**Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber.** 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113 (1): 151–84.

**Angrist, Joshua D., and William N. Evans.** 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88 (3): 450–77.

**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 Alan B. Krueger.** 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (4): 979–1014.

**Ashraf, Quamrul, and Oded Galor.** 2013a. "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review* 103 (1): 1–46.

**Ashraf, Quamrul, and Oded Galor.** 2013b. *Replication data for: The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development.* Inter-university Consortium for Political and Social Research [distributor], American Economic Association. https://doi.org/10.3886/E112588V1.

**Autor, David H., David Dorn, and Gordon H. Hanson.** 2013a. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (6): 2121–68.

**Autor, David H., David Dorn, and Gordon H. Hanson.** 2013b. *Replication Data for: The China Syndrome: Local Labor Market Effects of Import Competition in the United States.* Inter-university Consortium for Political and Social Research [distributor], American Economic Association. https://doi.org/10.3886/E112670V1.

**Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky.** Forthcoming. "When Is TSLS Actually LATE?" *Review of Economic Studies*.

**Chan, David C., David Card, and Lowell Taylor.** 2023. "Is There a VA Advantage? Evidence from Dually Eligible Veterans." *American Economic Review* 113 (11): 3003–43.

**Chetty, Raj, John N. Friedman, and Jonah E. Rockoff.** 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

**Chyn, Eric, Brigham Frandsen, and Emily C. Leslie.** 2025. "Examiner and Judge Designs in Economics: A Practitioner's Guide." *Journal of Economic Literature* 63 (2): 401–39.

**Cragg, John G.** 1983. "More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form." *Econometrica* 51 (3): 751–63.

**Danieli, Oren, Daniel Nevo, Itai Walk, Bar Weinstein, and Dan Zeltzer.** 2026. *Data and Code for: "Negative Control Falsification Tests for Instrumental Variable Designs."* American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E238658V1.

**Davies, Neil M., Kyla H. Thomas, Amy E. Taylor, Gemma M. J. Taylor, Richard M. Martin, Marcus R. Munafò, and Frank Windmeijer.** 2017. "How to Compare Instrumental Variable and Conventional Regression Analyses Using Negative Controls and Bias Plots." *International Journal of Epidemiology* 46 (6): 2067–77.

**Dawid, A. Philip.** 1979. "Conditional Independence in Statistical Theory." *Journal of the Royal Statistical Society: Series B (Methodological)* 41 (1): 1–15.

**De Giorgi, Giacomo, Anders Frederiksen, and Luigi Pistaferri.** 2020. "Consumption Network Effects." *Review of Economic Studies* 87 (1): 130–63.

**Deming, David J.** 2014a. "Using School Choice Lotteries to Test Measures of School Effectiveness." *American Economic Review* 104 (5): 406–11.

**Deming, David J.** 2014b. *Replication Data for: "Using School Choice Lotteries to Test Measures of School Effectiveness*." Inter-university Consortium for Political and Social Research [distributor], American Economic Association. https://doi.org/10.3886/E112805V1.

**Diegert, Paul, Matthew A. Masten, and Alexandre Poirier.** 2022. "Assessing Omitted Variable Bias when the Controls Are Endogenous." Preprint, arXiv. https://doi.org/10.48550/arXiv.2206.02303.

**Doyle, Joseph J., Jr., John A. Graves, Jonathan Gruber, and Samuel A. Kleiner.** 2015. "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns." *Journal of Political Economy* 123 (1): 170–214.

**Dukes, O., D. B. Richardson, Z. Shahn, J. M. Robins, and E. J. Tchetgen Tchetgen.** 2025. "Using Negative Controls to Identify Causal Effects with Invalid Instrumental Variables." *Biometrika* 112 (1): asae064.

**Eggers, Andrew C., Guadalupe Tuñón, and Allan Dafoe.** 2024. "Placebo Tests for Causal Inference." *American Journal of Political Science* 68 (3): 1106–21.

**Frandsen, Brigham, Lars Lefgren, and Emily Leslie.** 2023. "Judging Judge Fixed Effects." *American Economic Review* 113 (1): 253–77.

**Glymour, M. Maria, Eric J. Tchetgen Tchetgen, and James M. Robins.** 2012. "Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions." *American Journal of Epidemiology* 175 (4): 332–39.

**Guidetti, Bruna, Paula Pereda, and Edson Severnini.** 2021. "'Placebo Tests' for the Impacts of Air Pollution on Health: The Challenge of Limited Health Care Infrastructure." *AEA Papers and Proceedings* 111: 371–75.

**Hastie, Trevor J., and Robert J. Tibshirani.** 1990. *Generalized Additive Models*. Chapman and Hall/CRC.

**Hastings, Justine, and Ebonya Washington.** 2010. "The First of the Month Effect: Consumer Behavior and Store Responses." *American Economic Journal: Economic Policy* 2 (2): 142–62.

**Heinze-Deml, Christina, Jonas Peters, and Nicolai Meinshausen.** 2018. "Invariant Causal Prediction for Nonlinear Models." *Journal of Causal Inference* 6 (2): 20170016.

**Huber, Martin, and Giovanni Mellace.** 2015. "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints." *Review of Economics and Statistics* 97 (2): 398–411.

**Imbens, Guido W.** 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79.

**Jacob, Brian, Lars Lefgren, and Enrico Moretti.** 2007. "The Dynamics of Criminal Behavior: Evidence from Weather Shocks." *Journal of Human Resources* 42 (3): 489–527.

**Jäger, Simon, Jörg Heining, and Nathan Lazarus.** 2024. "How Substitutable Are Workers? Evidence from Worker Deaths." Unpublished.

**Keele, Luke, Qingyuan Zhao, Rachel R. Kelz, and Dylan Small.** 2019. "Falsification Tests for Instrumental Variable Designs with an Application to Tendency to Operate." *Medical Care* 57 (2): 167–71.

**Kitagawa, Toru.** 2015. "A Test for Instrument Validity." *Econometrica* 83 (5): 2043–63.

**Kling, Jeffrey R.** 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76.

**Li, Chun, and Xiaodan Fan.** 2020. "On Nonparametric Conditional Independence Tests for Continuous Variables." *WIREs Computational Statistics* 12 (3): e1489.

**Lipsitch, Marc, Eric Tchetgen Tchetgen, and Ted Cohen.** 2010. "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies." *Epidemiology* 21 (3): 383–88.

**Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott.** 2013. "Do Political Protests Matter? Evidence from the Tea Party Movement." *Quarterly Journal of Economics* 128 (4): 1633–85.

**Martin, Gregory J., and Ali Yurukoglu.** 2017. "Bias in Cable News: Persuasion and Polarization." *American Economic Review* 107 (9): 2565–99.

**Miao, Wang, Zhi Geng, and Eric J. Tchetgen Tchetgen.** 2018. "Identifying Causal Effects with Proxy Variables of an Unmeasured Confounder." *Biometrika* 105 (4): 987–93.

**Moretti, Enrico.** 2021. "The Effect of High-Tech Clusters on the Productivity of Top Inventors." *American Economic Review* 111 (10): 3328–75.

**Mourifié, Ismael, and Yuanyuan Wan.** 2017. "Testing Local Average Treatment Effect Assumptions." *Review of Economics and Statistics* 99 (2): 305–13.

**Nunn, Nathan, and Nancy Qian.** 2014a. "US Food Aid and Civil Conflict." *American Economic Review* 104 (6): 1630–66.

**Nunn, Nathan, and Nancy Qian.** 2014b. *Replication Data for: "US Food Aid and Civil Conflict.* Inter-university Consortium for Political and Social Research [distributor], American Economic Association. .

**Oster, Emily.** 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.

**Pearl, Judea.** 2009a. "Causal Inference in Statistics: An Overview." *Statistical Surveys* 3: 96–146.

**Pearl, Judea.** 2009b. *Causality*. Cambridge University Press.

**Pearl, Judea, and Azaria Paz.** 1986. "Graphoids: Graph-Based Logic for Reasoning about Relevance Relations or When Would X Tell You More About Y If You Already Know Z?" *ECAI'86: Proceedings of the 7th European Conference on Artificial Intelligence*, Vol. 2, edited by J. B. H. du Boulay, David Hogg, and Luc Steels, 357–63. North-Holland.

**Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt.** 2019. "Poorly Measured Confounders Are More Useful on the Left than on the Right." *Journal of Business & Economic Statistics* 37 (2): 205–16.

**Ramsey, James Bernard.** 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society: Series B (Methodological)* 31 (2): 350–71.

**Rosenzweig, Mark R., and Kenneth I. Wolpin.** 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38 (4): 827–74.

**Shah, Rajen D., and Jonas Peters.** 2020. "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure." *Annals of Statistics* 48 (3): 1514–38.

**Shi, Xu, Wang Miao, and Eric Tchetgen Tchetgen.** 2020. "A Selective Review of Negative Control Methods in Epidemiology." *Current Epidemiology Reports* 7 (4): 190–202.

**Tchetgen Tchetgen, Eric J., Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao.** 2024. "An Introduction to Proximal Causal Inference." *Statistical Science* 39 (3): 375–90.

**Textor, Johannes, Benito van der Zander, Mark S. Gilthorpe, Maciej Liskiewicz, and George T. H. Ellison.** 2016. "Robust Causal Inference Using Directed Acyclic Graphs: The R Package 'Dagitty.'" *International Journal of Epidemiology* 45 (6): 1887–94.

**Wood, Simon N.** 2006. *Generalized Additive Models*. Chapman and Hall/CRC.