THE PINCHAS SAPIR CENTER

FOR DEVELOPMENT

Tel Aviv University

המרכז לפיתוח ע"ש פנחס ספיר
ליד אוניברסיטת תל אביב
"עמותה רשומה"
580010221

THE PINHAS SAPIR CENTER FOR DEVELOPMENT
TEL AVIV UNIVERSITY

Negative Control Falsification Tests for Instrumental Variable Designs

Oren Danieli, Daniel Nevo, Itai Walk, Bar Weinstein, Dan Zeltzer

Discussion Paper No. 3-2025

Negative Control Falsification Tests for Instrumental Variable Designs*

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

Draft: April 9, 2025

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 that variables must satisfy in order to serve as negative controls. We show that these falsification tests examine not only independence and the exclusion restriction, but also functional form assumptions. Our analysis reveals that conventional applications of these tests may flag problems even in valid IV designs. We offer implementation guidance to address these issues.

The identification assumptions in instrumental variable (IV) designs cannot be directly tested. Instead, researchers often use indirect falsification, or "placebo," tests. Reviewing the most-cited papers in five leading economics journals, we find that 51% of IV studies employ such falsification tests. The large majority of falsification tests fall into two categories: 75% 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% 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 correlated with variables that resemble the IV but do not affect the treatment. Extensive literature has developed a theoretical framework for negative control falsification tests in classic causal settings (Lipsitch, Tchetgen Tchetgen and Cohen, 2010; Shi, Miao and Tchetgen Tchetgen, 2020), but not for the assumptions underlying IV designs.[1] This paper aims to fill this gap.

We propose that practitioners first identify potential threats to their IV validity, which we formally define. 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. The theory highlights two common pitfalls, which can falsely flag problems in

* 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, Technion, Hebrew University of Jerusalem, Tel Aviv University, EIEF Conference and SOLE 2024 meeting 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.

† All authors are from Tel Aviv University. Danieli is the corresponding author. Email: orendanieli@tauex.tau.ac.il.

[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.

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 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 higher viewership. One concern is that unobserved local conservativeness (U, the APO variable) affects not only the republication vote share (the outcome) but also the channel position (the IV), as marked with 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 (2014), who examine the effect of US food aid (X) on conflicts 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 with 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, the lagged

[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 II.A.

[3]Throughout the paper, we use directed acyclic graphs (DAGs) to visualize complex structures, as advocated by Imbens (2020). Appendix C outlines the theory presented in this paper within the formal causal DAG framework (Pearl, 2009)

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 correlate with channel position, it would imply that 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 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.

The theory highlights 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 correlated 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 correlated with the original IV and therefore will be associated with the outcome due to this correlation, 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). This means that orange production and conflict (Y) would be correlated 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 theory 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, Forthcoming). 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) recommends using such control variables as NCOs instead. Eggers, Tuñón and Dafoe (2023) discuss the usage of placebo tests in the social sciences more broadly. We contribute to this literature by outlining a theoretical framework for 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., 2024). We apply the theory of 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. Our work is related to Davies et al. (2017), who use negative controls in IV contexts without developing a theoretical framework for such an approach. We find several important differences in the theoretical framework of negative controls for assessing IVs compared to their usage in the standard treatment-outcome setting.

The rest of this paper proceeds as follows. Section I surveys the current practice of falsification tests for IV designs. Section II presents the theory 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. Appendix B provides additional details on the survey construction, and the results are summarized in Table 2.

We highlight five key findings from this survey. First, falsification tests are widely used in IV analyses. Approximately half (51%) 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 described in the intro-

duction and formalized in this paper. As outlined earlier, these tests can be divided into two types: *negative control outcome* (NCO) tests, which check for associations between the IV and variables it should not be associated with, and *negative control instrument* (NCI) tests, which examine associations between the outcome and variables it should not be associated with. Among surveyed papers using falsification tests, 75% used NCO tests (Column 3) and 24% used NCI tests (Column 4). All other types of falsification tests combined were used in 21% of the papers (Column 5); Appendix B 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% of all NCO tests. The remaining NCO tests often follow a similar logic.[4] 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% of the papers surveyed 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% of of the papers using only one. This finding suggests that researchers use only a subset of the available relevant negative controls. As we demonstrate in Section III, the theory 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. Theory of Negative Controls in IV Settings

In this section, we present the theory of negative control tests for IV designs. The theory is constructed using the terminology of potential outcomes, and we use DAGs for examples and intuition. Appendix C introduces the basic relevant concepts for DAG theory and replicates the key definitions and theorems using DAGs.

A. The Threats to the Identification

IV 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.[5] 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.

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

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

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 more general independence assumption (e.g., Abadie, 2003). Here, 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 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, i.e., $Y(x) = Y(Z, x)$.

Assumption 2 (Exclusion restriction)**.** *For all z, x, $Y(z, x) = Y(x)$.*

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 who have 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 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[6]

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

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, additional assumptions are also necessary. For example, the design of Angrist, Imbens and Rubin (1996) also requires treatment independence,

[6]When the exclusion restriction does not hold, $Y(x)$ is still properly defined but does not equal $Y(z, x)$ for every value of z. Therefore, recalling that $Y(x) = Y(Z, x)$, violation of the exclusion restriction implies $Z \not\perp\!\!\!\perp Y(x)$.

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. Appendix D 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 *alternative path outcome* (APO) variables. These variables are presumably associated with the outcome. The threat is that they are 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.*

1) Latent IV validity. $Z \perp\!\!\!\perp Y(x)|U$.

2) 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 make the IV and the potential outcome conditionally independent. Using again the example of Martin and Yurukoglu (2017), the share of Republican votes in 1996 is only an imperfect 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 Appendix E.1). Path indication also rules out variables that could be related to both the IV and the outcome without generating a correlation between them. For example, this would occur if a variable is correlated with the outcome for some subpopulation but potentially correlated with the IV only for a separate subpopulation. Two examples are provided in Appendix E.2 and Appendix E.3.

The second type of alternative path variables is *alternative path instrument* (API) variables. API variables are known to be associated with the IV, and the concern is an association they might have with the outcome. This is in contrast to APO variables, which are known to be associated with the outcome, and the concern is their possible association with the IV. Using the previous example of Nunn and Qian (2014), 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.*

1) Latent IV validity. $Z \perp\!\!\!\perp Y(x)|U$.

2) 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 $Z \perp\!\!\!\perp Y(x)$, 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 Appendix E.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.*

1) The NCO assumption. $NC \perp\!\!\!\perp Z|U$.

2) U-comparability. $NC \not\perp\!\!\!\perp U$.

The NCO assumption guarantees that any path between the IV and the NCO must go through the APO variable U. It rules out variables that have other 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 assignment (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 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 non-respiratory hospital admissions, a seemingly unrelated outcome, as NCOs. One might expect that these admissions would only correlate with flawed IVs for air pollution. However, Guidetti, Pereda and Severnini (2021) demonstrate otherwise. They find that air pollution increases non-respiratory admissions through hospital congestion caused by a surge in respiratory admissions. Therefore, non-respiratory admissions are not informative about the IV validity as they correlate with both flawed and valid IVs. Formally, non-respiratory admissions correlate with the IV, not through any APO variable but due to the unrelated mechanism of congestion. Hence, non-respiratory 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 correlated 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.*

1) The NCI assumption. $NC \perp\!\!\!\perp Y|Z,U$.

2) U-comparability $NC \not\perp\!\!\!\perp U|Z$.

The NCI assumption guarantees that any path between the outcome and the NCI goes through the API variable U or the IV Z. It rules out variables that have other 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 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 (2014), 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 production (Z) and, therefore, would be correlated 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. Moreover, in cases where the IV effect on the outcome is heterogeneous, any variable associated with the source of heterogeneity cannot serve as an NCI. 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 Appendix D, 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 should equal zero. Indeed, they found no evidence to reject this hypothesis, which supports their design validity.

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 Appendix D. The appendix proof covers a more general version of Theorem 1 for designs that include control variables (discussed in Section II.D). 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 ($Z \not\perp\!\!\!\perp Y(x)$); i.e., 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 (2014) regress their outcome, conflicts, on various alternative crop production (e.g., oranges), which are the NCIs, controlling for the original IV, wheat production. They are unable to 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 (2014), orange production (the NCI) is associated with conflicts (the outcome), as both are associated 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.*

Appendix D 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 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.[7] 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.

As a result, unconditional independence tests between a negative control and the outcome are unique to IV settings. In non-IV settings, there is no exclusion restriction, and therefore, independence tests between a negative control and the outcome, carried out to detect unmeasured confounding, are always done conditionally.[8]

Nevertheless, researchers can choose to always control for the IV. Since both NC and Z are observed, the condition $NC \perp\!\!\!\perp Z$ can be empirically tested. However, researchers might opt to skip this test and condition on the IV anyway. In a linear model, adding an additional control that is uncorrelated with 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 $Z \perp\!\!\!\perp Y(x) | C$. Appendix D presents the theory of negative controls when control variables are included.

When the IV is presumably valid only conditional on a vector of control variables C, an NCO

[7] The NCI assumption is that payday timing only correlates with crime only through the timing of economic activity.
[8] The analog of NCI in non-IV settings is negative control exposure (NCE). NCE tests always condition on the exposure.

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.[9] Blandhol et al. (2022) show that 2SLS requires the following linearity assumption to satisfy their definition of a *weakly causal estimand*.[10]

Assumption 3 (Rich covariates)**.** *The conditional expectation of the IV is linear in the control specification. Namely,* $\mathbb{E}[Z|C] = \gamma_C' C$*, for some vector* γ_C*.*

Combining the null hypothesis of NCO tests (2) and rich covariates, we expect that

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

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 formalizes this argument.

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

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

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.*

In many cases, researchers 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*

$$\beta = (\beta_Z, \beta_C') = \underset{b_Z, b_C}{\arg\min}\, \mathbb{E}[NC - b_Z Z - b_C' C]^2$$

[9] 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.

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

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, $\mathbb{E}[Y|Z, C] = \theta_Z Z + \theta_C' C$, for some θ_Z and vector θ_C.*

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

$$\mathbb{E}[Y|Z, C, NC] = \theta_Z Z + \theta_C' C. \tag{5}$$

This equation also provides a more specific null hypothesis, which can be tested with OLS. The following corollary 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*

$$\theta = (\theta_Z, \theta_C', \theta_{NC}) = \underset{b_Z, b_c, b_{NC}}{\arg\min}\, \mathbb{E}[Y - b_Z Z - b_C' C - b_{NC} NC]^2$$

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., 2022). For NCI tests, the null can be rejected because the CSRF assumption is violated. However, unlike rich covariates, CSRF is not a necessary assumption for 2SLS analysis, implying that 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 A1. 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 III.A. 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 III.B.

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 III.C 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 III.D.

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 (2013) and Deming (2014) to discuss NCO tests and Ashraf and Galor (2013) and Nunn and Qian (2014) to discuss NCI tests.[11] Appendix Table A1 summarizes the IV designs in these papers and the negative controls they used in their falsification tests. Appendix F 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 this 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, 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

[11] Ashraf and Galor (2013) and Nunn and Qian (2014) are the two most cited AER papers published since 2013 that use an NCI test. Similarly, Autor, Dorn and Hanson (2013) is the most cited AER paper published after 2013 that uses an NCO test. Deming (2014) was selected to demonstrate how our proposed follow-up analysis can be used to diagnose and correct problems with the IV design in Section III.D.

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 (2014), 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. In Figure 1, panels A and B 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 search for NCOs and test their association with the IV. For API variables, researchers should search for NCIs and test their association with the outcome, conditionally on the IV. The type of violation (outcome independence or exclusion restriction) can be useful for thinking of 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 theoretical framework.

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.[12] If the IV is associated with a predetermined

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

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 (2014), 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 IV with the predetermined variables. In particular, we found that the construction of the IV involved non-random components that require additional controls; see Section III.D.

Predetermined variables are also useful NCOs when the IV is not quasi-random. For example, Autor, Dorn and Hanson (2013) 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 non-random, it might be associated with other labor market conditions, which in turn could be associated with the outcome (e.g., Chinese imports are 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, i.e., 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 such unexpected shocks should not be autocorrelated, leads and lags of the variable used as the IV can serve as NCOs. For example, Jäger, Heining and Lazarus (Forthcoming) 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 non-random 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 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 should not be autocorrelated. 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 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 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 previously discussed example of Angrist and Evans (1998), the concern is that same-sex sibship IV may affect female labor supply due to hand-me-downs, thus forming an alternative path. To evaluate this concern, Rosenzweig and Wolpin (2000) check the correlation of same-sex sibship and an alternative outcome—clothing expenditure.[13]

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 correlated with the API variable. For example, as previously discussed (and illustrated in Panel C of Figure 1), Nunn and Qian (2014) 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. Similarly, Ashraf and Galor (2013) replace their original IV, distance from Addis-Ababa, with distance from London, Tokyo, and Mexico City.

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 correlated 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. For example, Moretti (2021) studies the effect of the size of high-tech clusters on productivity. As an IV, he uses predicted cluster size based on the expansion of local firms outside the cluster. Moretti then ascertains that future predicted cluster size is also not correlated with productivity. This relies on the fact that IV leads, which are based on events that occur after the outcome, cannot influence it. IV leads share similarities with the IV and are therefore likely to be associated with the API variable. To satisfy the NCI assumption, the outcome must not influence future realizations of the IV. In the example of Moretti (2021), regional productivity cannot affect the expansion of local firms in other locations.

When practitioners observe IV leads, they need to consider whether they expect the IVs to be autocorrelated. When the IVs are expected to be autocorrelated (as in Moretti, 2021) an NCI

[13] In this example, the NCO is the APO variable itself, so the NCO assumption is trivially satisfied (see Section II.B)

strategy can be used. When the IVs are presumably uncorrelated, and 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. This approach does not require full knowledge of the API variable. 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 any factors that influence a child's quarter of birth and are suspected to affect wages (the outcome).

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 act as a *collider*; Pearl, 2009). 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. This, in turn, implies that the API variable affects the outcome, violating outcome independence.

To satisfy the NCI assumption, these NCIs should have no association with the outcome other than through the IV.[14] 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.

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 et al. use traffic patterns, a proxy for economic activity, as an NCI (NC_4). Testing whether traffic patterns are correlated with crime, conditional on lagged weather, constitutes an NCI test for this alternative path. If a correlation exists, 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, that could be other factors that displace economic activity (e.g., payday schedule; see the discussion of this example in Section II.C).

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

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 in NCI variables that are intentionally similar to the original IV. Such variables are often strongly correlated with the IV but only weakly correlated with the API variable conditional on the IV.

C. Choosing a Statistical Test

As discussed in Section II.C, 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, 2023).

When multiple negative controls are available, Model (6) can be estimated separately for each negative control. However, correcting for multiple hypotheses is necessary, which reduces statistical power. 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.[15] An F-test can be used to evaluate the null hypothesis $H_0 : \gamma'_{NC} = 0'$ under standard assumptions.[16]

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 II.D, the CSRF assumption is not necessary for causal identification and can be violated even under random assignment.[17] Therefore, tests based on Model (8) may reject 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 semi-parametric or non-parametric 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. Moreover, these tests only test the mean independence of the IV with the NCO or the outcome with the NCI. In some cases, researchers should use more general tests.

Researchers should consider replacing the functional form assumptions in two cases. First, tests with more flexible functional forms can help relax undesirable functional form assumptions. In particular, when using an NCI, researchers should test for non-linear associations whenever possible to avoid testing the CSRF assumption, which is unnecessary. 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 in such contexts.

Second, more flexible tests are useful when researchers suspect a non-linear association between the NCO and the IV or between the NCI and the outcome. Such non-linear 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, 2014), 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 non-monotonic relationship between an average temperature NCI and the outcome.

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

[16] 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.

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

To estimate more complex functional forms, researchers can include higher-order polynomial terms or interactions in regression models. Alternatively, they can use semi- or non-parametric 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. For example, IV quantile regression (e.g., Chernozhukov and Hansen, 2008) relies on the broader notion of conditional independence. In such cases, researchers can use quantile regression of the IV on the NCO or the outcome on the NCI, with appropriate controls. A variety of other conditional independence tests can also be considered (see Heinze-Deml, Peters and Meinshausen, 2018; 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.[18] 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 3 presents the NCO tests results using data from Autor, Dorn and Hanson (2013) and Deming (2014). Column (2) shows that a test based on a single NCO fails to reject the null hypothesis in both cases.[19] However, Column (3) demonstrates that using multiple NCOs and applying a Bonferroni correction leads to rejection of the null, as does a joint F-test (Column 4). These results underscore how theory-guided inclusion of additional NCOs enhances test power.

To examine non-linear associations, we implement the common semi-parametric 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 and 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.[20]

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 (5) and (6) in Table 3 implement a GAM test with and without assuming linearity in the controls. As expected, the GAM test underperforms in smaller samples.

Table 4 presents the NCI test results for Nunn and Qian (2014) and Ashraf and Galor (2013). 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 (2014), failure to condition leads to rejection of the null, suggesting a false rejection due to the

[18] 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.

[19] For Autor, Dorn and Hanson (2013), 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.

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

path between the NCI and the outcome through the IV. For Ashraf and Galor (2013), the null is not rejected in either case, likely due to limited statistical power. Columns (3) and (4) of Table 4 implement multiple separate NCI tests with Bonferroni corrections and joint F-tests using all NCIs. In both approaches, the null is not rejected.[21]

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 non-parametric 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 diagnostic scatter plot of the correlation of each negative control with the IV against its correlation with the outcome can highlight potential alternative pathways. Appendix Figure A2 displays such diagnostics for Deming (2014). Deming (2014) 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 (see Appendix Table A1 for details). This IV satisfies outcome independence only when controlling for the relevant school value-added measures and the probability of winning the lotteries. The scatter plot reveals that the NCO with the strongest 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 satisfied conditional only on the winning probabilities (i.e., the schools the student applied to).[22]

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 the strength of their correlations with the IV or outcome depends on the strength of their correlation with the alternative path variables. Thus, weak correlations between a negative control and the IV or outcome might still mask strong alternative paths.

Non-Rejection 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 found to be uncorrelated with students' neighborhoods could still be correlated with students' abilities within neighborhoods. Second, an invalid IV design may pass the test due to limited statistical power.

[21] 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 (2014), we estimate a GAM model assuming linear controls; see Appendix F.3.

[22] While estimates using the original lottery as the IV are noisier, we cannot reject the main conclusions.

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 correlated 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 we discussed in Section II.D. Moreover, including NCIs in the reduced form will only decrease precision because they are correlated with the IV and not with the outcome.[23]

REFERENCES

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics*, 113(2): 231–263.

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): 151–184.

Angrist, Joshua D., and Alan B Krueger. 1991. "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014.

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): 450–477.

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): 444–455.

[23] 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.

Ashraf, Quamrul, and Oded Galor. 2013. "The 'Out of Africa' hypothesis, human genetic diversity, and comparative economic development." *American Economic Review*, 103(1): 1–46.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China syndrome: Local labor market effects of import competition in the United States." *American Economic Review*, 103(6): 2121–2168.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When is TSLS actually LATE?" *NBER Working Paper 29709*.

Chan, David C., David Card, and Lowell Taylor. 2023. "Is there a VA advantage? Evidence from dually eligible veterans." *American Economic Review*, 113(11): 3003–3043.

Chernozhukov, Victor, and Christian Hansen. 2008. "Instrumental variable quantile regression: A robust inference approach." *Journal of Econometrics*, 142(1): 379–398.

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): 2593–2632.

Chyn, Eric, Brigham Frandsen, and Emily C Leslie. Forthcoming. "Examiner and Judge Designs in Economics: A Practitioner's Guide." *Journal of Economic Literature*.

Cragg, John G. 1983. "More efficient estimation in the presence of heteroscedasticity of unknown form." *Econometrica: Journal of the Econometric Society*, 751–763.

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): 2067–2077.

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): 130–163.

Deming, David J. 2014. "Using school choice lotteries to test measures of school effectiveness." *American Economic Review*, 104(5): 406–11.

Diegert, Paul, Matthew A. Masten, and Alexandre Poirier. 2022. "Assessing omitted variable bias when the controls are endogenous." *arXiv preprint arXiv:2206.02303*.

Doyle, Joseph J., 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, DB Richardson, Z Shahn, JM Robins, and EJ Tchetgen Tchetgen. 2024. "Using negative controls to identify causal effects with invalid instrumental variables." *Biometrika*, asae064.

Eggers, Andrew C., Guadalupe Tuñón, and Allan Dafoe. 2023. "Placebo tests for causal inference." *American Journal of Political Science*.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging judge fixed effects." *American Economic Review*, 113(1): 253–277.

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): 332–339.

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, and Robert Tibshirani. 1990. *Generalized Additive Models.* Vol. 43 of *Monographs on Statistics and Applied Probability*, New York: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): 1–35.

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. Forthcoming. "How substitutable are workers? Evidence from worker deaths." *American Economic Review*.

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): 167–171.

Kitagawa, Toru. 2015. "A test for instrument validity." *Econometrica*, 83(5): 2043–2063.

Kling, Jeffrey R. 2006. "Incarceration length, employment, and earnings." *American Economic Review*, 96(3): 863–876.

Li, Chun, and Xiaodan Fan. 2020. "On nonparametric conditional independence tests for continuous variables." *Wiley Interdisciplinary Reviews: 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 (Cambridge, Mass.)*, 21(3): 383.

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): 1633–1685.

Martin, Gregory J., and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review*, 107(9): 2565–2599.

Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. "Identifying causal effects with proxy variables of an unmeasured confounder." *Biometrika*, 105(4): 987–993.

Moretti, Enrico. 2021. "The effect of high-tech clusters on the productivity of top inventors." *American Economic Review*, 111(10): 3328–3375.

Mourifié, Ismael, and Yuanyuan Wan. 2017. "Testing local average treatment effect assumptions." *Review of Economics and Statistics*, 99(2): 305–313.

Nunn, Nathan, and Nancy Qian. 2014. "US food aid and civil conflict." *American Economic Review*, 104(6): 1630–1666.

Oster, Emily. 2019. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

Pearl, Judea. 2009. *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?" 357–363. North-Holland. PDF available.

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): 205–216.

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): 350–371.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'natural experiments' in economics." *Journal of Economic Literature*, 38(4): 827–874.

Shah, Rajen D., and Jonas Peters. 2020. "The hardness of conditional independence testing and the generalised covariance measure." *Annals of Statistics*, 48(3): 1514–1538.

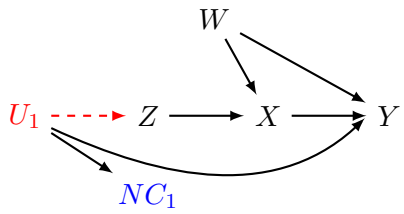
Shi, Xu, Wang Miao, and Eric Tchetgen Tchetgen. 2020. "A selective review of negative control methods in epidemiology." *Current Epidemiology Reports*, 7(4): 190–202.

Tchetgen Tchetgen, Eric J, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. 2024. "An introduction to proximal causal inference." *Statistical Science*, 39(3): 375–390.

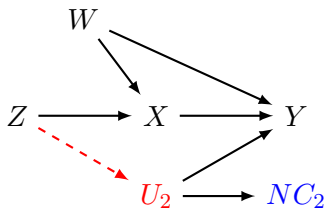
Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Figure 1. Negative Control Falsification Tests: Graphical Illustrations

Negative Control Outcome Tests

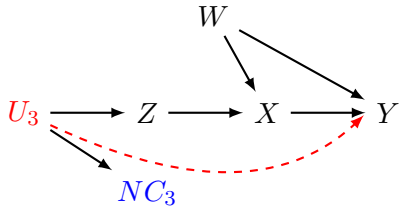


(A) $NC_1 \not\perp\!\!\!\perp Z$ implies that the dashed arrow exists, thus violating the outcome independence.

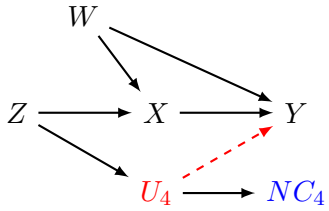


(B) $NC_2 \not\perp\!\!\!\perp Z$ implies that the dashed arrow exists, thus violating the exclusion restriction.

Negative Control Instrument Tests



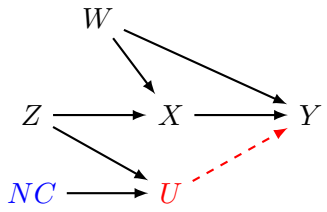
(C) $NC_3 \not\perp\!\!\!\perp Y|Z$ implies that the dashed arrow exists, thus violating the outcome independence.



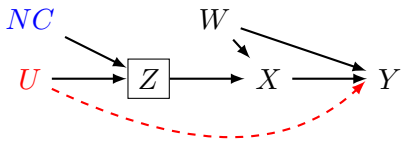
(D) $NC_4 \not\perp\!\!\!\perp Y|Z$ implies that the dashed arrow exists, thus violating the exclusion restriction.

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 the outcome independence (Panel A) or exclusion restriction (Panel B) assumptions. 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 the outcome independence (Panel C) or exclusion restriction (Panel D) assumptions. An observed negative control instrument (NC_3 or NC_4) related to each API variable can examine these concerns by testing whether $NC_i \perp\!\!\!\perp Y|Z$.

Figure 2. Illustration of Scenarios Related to Negative Control Instrument Tests



(A) An unconditional NCI test



(B) NCI causally affecting the IV

Notes: Each panel represents a different scenario related to negative control instrument (NCI) 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 $U \not\perp\!\!\!\perp Y|Z$.

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

A. Negative Control Outcome Tests

Paper	Treatment	Outcome	IV	Threat	NCO (IV should not be correlated with it)
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 (2013)	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 (Forthcoming)	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

B. Negative Control Instrument Tests

Paper	Treatment	Outcome	IV	Threat	NCI (conditional on IV, outcome should not be correlated with it)
Nunn and Qian (2014)	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 (2013)	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 permutated 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

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

			Falsification Test Characteristics (Share of Papers that Included Falsification Tests)						
			Type of Test			Test Specification			
	Papers Reviewed	Share with Falsification Tests	Negative Control Outcome	Negative Control Instrument	Other	Replace Outcome with NCO	Replace IV with NCI	Add NCI to reduced-form	# Negative Controls Used (median)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(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 (IV) 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. Appendix B provides additional details on the survey construction.

Table 3—Illustrative Applications of Negative Control Outcome Tests

	Original Analysis	**Alternative Analyses**					
Type of Test:	Single linear model without controls	Single linear model	Multiple linear models	F-Test	GAM, Linear Controls	GAM, Nonlinear Controls	Number of Observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Autor, Dorn and Hanson (2013)	0.004	0.593	<0.001	<0.001	1.000	1.000	722
Deming (2014)	-	0.686	<0.001	<0.001	<0.001	<0.001	2343

Notes: This table presents p-values from different NCO tests using data from Autor, Dorn and Hanson (2013) and Deming (2014). Column (1) replicates one of the original falsification analyses, in which Autor, Dorn and Hanson replaced the outcome with the NCO in the same 2SLS specification as their main analysis (ibid., Table 2, Part II). Deming conducted no falsification tests. Columns 2–6 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 single test using one NCO, where the outcome is replaced with the NCO in the reduced form regression. For Autor, Dorn and Hanson the single NCO is the lagged outcome (in 1970), which is the same NCO reported in Column (1); for Deming this is lagged test scores (2002). Column (3) presents a Bonferroni-corrected p-value for multiple tests using all the NCOs, with the same specification as Column (2). Column (4) uses an F-test (Model (7)) with all NCOs jointly. Columns (5) and (6) use GAM tests with linear and smoothed controls, respectively (Model (9)).

Table 4—Illustrative Applications of Negative Control Instrument Tests

	Without Conditioning on the IV	**With Conditioning on the IV**			
	Single linear model	Single linear model	Multiple linear models	F-Test	Number of Observations
	(1)	(2)	(3)	(4)	(5)
Nunn and Qian (2014)	0.007	0.123	0.636	0.138	4572
Ashraf and Galor (2013)	0.234	0.778	1.000	0.994	145

Notes: This table presents p-values from different NCI tests using data from Nunn and Qian (2014) and Ashraf and Galor (2013), applying their original sets of NCIs (three and ten NCIs, respectively). Column (1) shows a single linear NCI test that, inappropriately, does not condition on the IV. The NCI with the lowest p-value is shown (grape production for Nunn and Qian and distance from Mexico City for Ashraf and Galor). Columns (2)–(4) condition on the IV: Column (2) implements a proper linear NCI test (Model (8)) using the same NCI as Column (1); Column (3) applies Bonferroni correction for multiple linear NCI tests; Column (4) uses an F-test for all NCIs jointly.

Online Supplementary Appendices

Negative Control Falsification Tests for Instrumental Variable Designs

Oren Danieli (orendanieli@tauex.tau.ac.il), Daniel Nevo, Itai Walk, Bar Weinstein, Dan Zeltzer

April 9, 2025

Appendix A. Additional Tables and Figures

Appendix Table A1—Papers Used to Illustrate Applications of Negative Control Tests

	Original Variable Description				# of Negative Controls	
	Outcome	Treatment	IV	Negative Controls	Original Paper	Our Analyses
	(1)	(2)	(3)	(4)	(5)	(6)
A. Negative Control Outcome Tests						
Autor, Dorn and Hanson (2013)	Local labor market outcomes	Import competition from China	Other countries' import competition from China	Lagged local labor market manufacturing employment	3	52
Deming (2014)	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 (2013)	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 (2014)	Level of civil conflict	Receipt of U.S. food aid	Variation in U.S. wheat production	Variation in U.S. 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 3 and Table 4. 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 3 and Table 4, including the original negative controls, as well as additional valid negative controls we found in the original data.

Appendix Figure A1. Steps for Implementing Negative Control Tests

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 II.A 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 II.B discuss the assumptions these proxies need to satisfy. Examples are discussed in Section III.B.

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:**

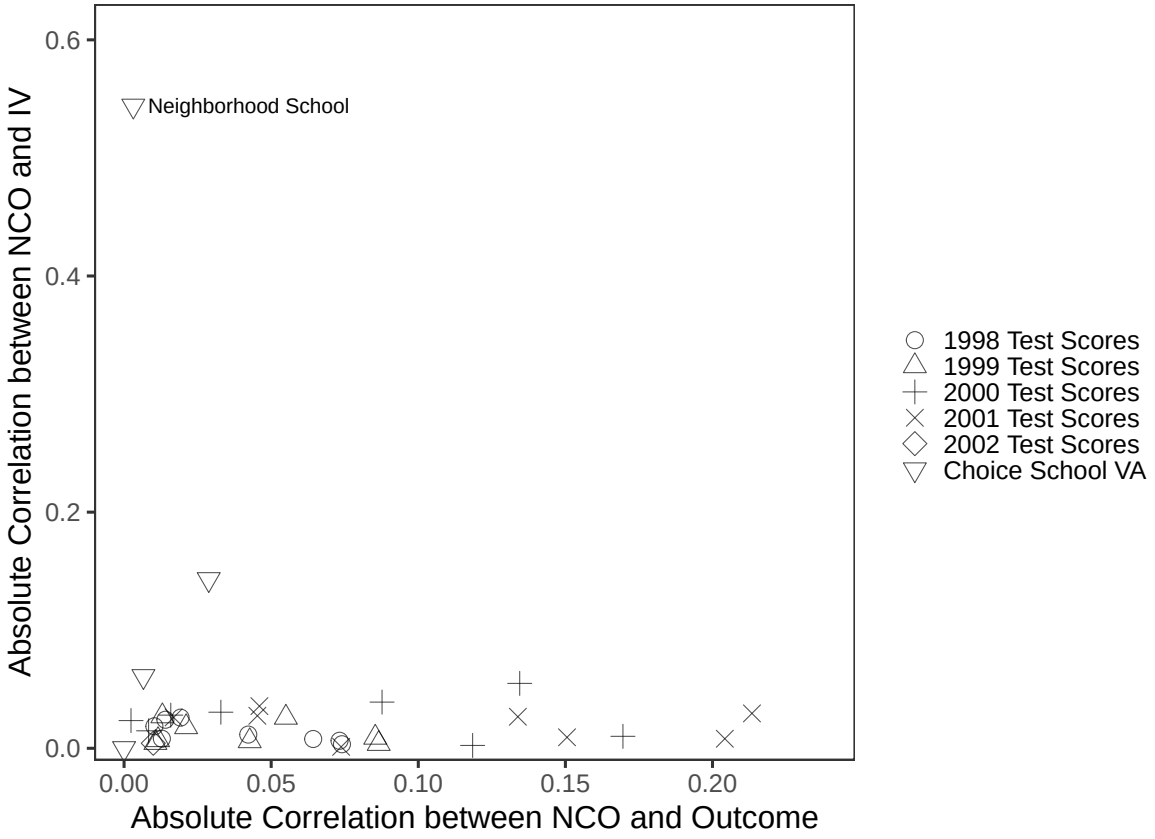
   If the null is rejected:

   - 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 A2. Correlations of NCOs with the IV and the Outcome in Deming (2014)



This figure shows a scatter plot of the absolute value of the correlation of different NCOs with the IV (on the y-axis) and the outcome (on the x-axis) in Deming (2014). 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 1st, 2nd, and 3rd submitted preferences, as well as their neighborhood school's VA (labeled Neighborhood School). See Section Appendix F.2 for details.

Appendix B. Details on Survey of Common Practices

Sample Construction. We used Google Scholar in November 2023 to assemble the list of relevant papers. We searched the terms "instrumental variable," "instrument," "2SLS," and "IV." We restricted the sample to articles with over 300 citations or, if published after 2020, over 100 citations. We examined all articles satisfying these criteria, published in five top-ranked economics journals: Review of Economic Studies, American Economic Review, Journal of Political Economy, Quarterly Journal of Economics, and Econometrica. Overall, our survey includes 140 papers.

We then searched the papers for strings related to falsification testing. This included "falsification," "negative control," "balance," "balancing," "valid," and "validity." Papers that did not include any of these strings were marked as not having any falsification test. We manually coded the type of falsification test for papers that included one of these strings. The results are summarized in Table 2 and discussed in Section I.

Other Falsification Tests. As discussed in Section I, we categorized all falsification tests used in surveyed papers into NCO tests, NCI tests, and other falsification tests. Other falsification tests include the following: negative control tests in non-IV settings, which examine only the first or second stage in a 2SLS estimation; "placebo population" analyses (Eggers, Tuñón and Dafoe, 2023; Glymour, Tchetgen Tchetgen and Robins, 2012; Keele et al., 2019), which involve repeating the analysis using a different population where the IV is not expected to affect the outcome; validating that the results are robust to including additional control variables; and using an over-identification test when more than one IV is available.

Appendix C. Formalization of IV Negative Control Tests Using DAG Theory

In this section, we present an alternative formalization of the theory of negative controls for IV designs using the language of DAGs. We first summarize fundamental concepts from DAG theory.[24] The DAGs we use intuitively throughout the paper can be formalized using the theory reviewed in this section.

1. Background DAG Theory

A directed graph is a set of nodes and directed edges. A *directed path* on a graph between two nodes, X_1 and X_2, is a sequence of edges, such that the first edge starts at X_1, each edge starts at the arrowhead of the former edge, and the last edge ends at X_2. A directed *acyclic* graph (DAG) contains no cycles; namely, there are no directed paths from a node to itself. While DAGs can be used to represent assumptions on joint probability functions without notions of causality, here we

[24] This is not an exhaustive overview of DAG use for causality, but only the elements necessary for our purposes. See Pearl (2009) for detailed presentation and theory.

interpret all DAGs as causal, such that the arrows represent a causal relationship. Consider a DAG denoted by G. The set of *parents* of node X_j, denoted by PA_j, is the set of all nodes with direct arrows into X_j. The *descendants* of X_j are all nodes with a directed path of any number of edges (including a single edge) from X_j to those variables; these are variables causally affected by X_j, directly or indirectly.

DAGs can be used to encode conditional independence assumptions on the joint distribution of variables represented by nodes in the DAG. Each DAG represents an infinite number of probability functions sharing the same conditional independence structure. The joint distribution of all variables in the DAG is the product of the conditional probability function of each variable X_j given its parents PA_j.[25] Formally, for M variables $X_1, ...X_M$, this factorization is

$$P(x_1, ..., x_M) = \prod_{j=1}^{M} P(x_j|PA_j),$$

where lower cases are realizations. Any probability function P that admits to the above factorization is said to be compatible with a DAG G.

A key result of DAG theory is the translation of the structure represented by a DAG into conditional independence conditions. This translation relies on the concept of d-separation, a graphical condition.

Definition A1 (d-separation (Pearl, 2009))**.** *A path p from node X_1 to node X_2 in graph G is said to be blocked by a set of nodes $\mathcal{A}$ if and only if*

1) *The path p contains a chain $X_1 \to A \to X_2$ or a fork $X_1 \leftarrow A \to X_2$ such that the middle node A is in $\mathcal{A}$, or*

2) *The path p contains a collider structure $X_1 \to B \leftarrow X_2$ such that the middle node, the collider B, is not in $\mathcal{A}$ and such that no descendant of B is in $\mathcal{A}$.*

The set $\mathcal{A}$ is said to d-separate node X_1 from node X_2 if and only if $\mathcal{A}$ blocks every path between X_1 and X_2. In this case, we write

$$(X_1 \perp\!\!\!\perp X_2|\mathcal{A})_G.$$

Building on the above structure, the following theorem states the direct implication of d-separation, a graphical condition, on conditional independence, a probabilistic statement.

Theorem A1 (Pearl 2009)**.** *Let X_1, X_2 be two variables that are d-separated by a set of variables $\mathcal{A}$ in graph G, $(X_1 \perp\!\!\!\perp X_2|\mathcal{A})_G$. Then,*

$$X_1 \perp\!\!\!\perp X_2|\mathcal{A}$$

[25] This assumption is called the Markov property.

in all probability functions compatible with G.[26]

IVs can also be defined using a graphical criterion (Pearl, 2009, Definition 7.4.1). Consider a DAG G with nodes Z, X, Y (and possibly additional nodes, as in the DAGs in Figure 1). We follow Pearl (2009) by using $G_{\overline{X}}$ to denote the version of the DAG G with all arrows entering X removed.

Definition A2. *A variable Z is an IV for treatment X on outcome Y if*

1) $(Z \perp\!\!\!\perp Y)_{G_{\overline{X}}}$

2) $(Z \not\perp\!\!\!\perp X)_{G}$.

The first condition corresponds to the condition stated in (1). It implies that there is no alternative path between the IV and the outcome other than through the treatment. The second condition corresponds to the IV relevance assumption. The negative control falsification tests for IVs we discuss in this paper focus on the first condition.

2. Negative Controls for IVs Using DAGs

We are now ready to present the theory of negative controls for IVs using DAGs. To this end, we begin by providing the definitions of APO and API variables and those of NCOs and NCIs. Then, we provide proofs of Theorems 1 and 2 using DAG theory.

We start with the graphical DAG-based definition of an APO variable.

Definition A3 (Alternative path outcome variable)**.** *A random variable U is an APO variable if the following two conditions hold.*

1) Latent IV validity. $(Z \perp\!\!\!\perp Y|U)_{G_{\overline{X}}}$.

2) Path indication. If $(Z \perp\!\!\!\perp Y)_{G_{\overline{X}}}$ *then* $(Z \perp\!\!\!\perp U)_{G}$.

This definition is similar to the definition of APO variables using potential outcomes (Definition 1). Latent IV validity implies that Z and Y are d-separated by U (excluding the path through X). Path indication implies that if the IV is valid, then there is no unblocked path between Z and U.

Similarly, we can define graphically an API variable.

Definition A4 (Alternative path instrument variable)**.** *A random variable U is an API variable if the following two conditions hold.*

1) Latent IV validity. $(Z \perp\!\!\!\perp Y|U)_{G_{\overline{X}}}$.

[26] Often $X_1 \perp\!\!\!\perp X_2|\mathcal{A}$ is written as $(X_1 \perp\!\!\!\perp X_2|\mathcal{A})_P$. We omit the P subscript for simplicity.

2) Path indication. If $(Z \perp\!\!\!\perp Y)_{G_{\overline{X}}}$ *then* $(U \perp\!\!\!\perp Y|Z)_G$.

Turning to negative control variables, the definitions of NCO and NCI are similar to the definitions using potential outcomes (Definitions 3 and 4).

Definition A5. *A random variable NC is an NCO if there exists an APO variable U such that the following two conditions hold.*

1) The NCO assumption. $(NC \perp\!\!\!\perp Z|\ U)_G$.

2) U-comparability. $(NC \not\perp\!\!\!\perp U)_G$.

Definition A6. *A random variable NC is an NCI if there exists an API variable U such that the following two conditions hold.*

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

2) U-comparability. $(NC \not\perp\!\!\!\perp U|Z)_G$.

We now provide a proof of Theorem 1 under the above DAG definitions. Following the condition of Theorem 1, the NCO test finds that $NC \not\perp\!\!\!\perp Z$. By Definition A5 we have that $(NC \perp\!\!\!\perp Z|\ U)_G$. From the test we know that $NC \not\perp\!\!\!\perp Z$, which implies $(NC \not\perp\!\!\!\perp Z)_G$ by the contrapositive of Theorem A1. Because $(NC \not\perp\!\!\!\perp Z)_G$, there is at least one open (unblocked) path between Z and NC. However, because $(NC \perp\!\!\!\perp Z|U)_G$, this path or paths are blocked by U. By Definition A1, this means that U is either in the middle of a chain or a fork on the open path between Z and NC. Thus, there is an unblocked path between Z and U, i.e., $(Z \not\perp\!\!\!\perp U)_G$ which, by path indication, implies that $(Z \not\perp\!\!\!\perp Y)_{G_{\overline{X}}}$, violating the first IV condition in Definition A2.

We turn to the proof of Theorem 2 under the DAG definitions. Following the condition of Theorem 2, the NCI test finds that $NC \not\perp\!\!\!\perp Y|Z$. By Definition A6 we have that $(NC \perp\!\!\!\perp Y|\ Z, U)_G$. From the test, we know that $NC \not\perp\!\!\!\perp Y|Z$, which implies $(NC \not\perp\!\!\!\perp Y|Z)_G$ by the contrapositive of Theorem A1. Because $(NC \not\perp\!\!\!\perp Y|Z)_G$, there is at least one open path between NC and Y, which is not blocked by Z. However, because $(NC \perp\!\!\!\perp Y|U, Z)_G$, this path or paths are blocked by U. By Definition A1, this means that U is either in the middle of a chain or a fork on the open path between Y and NC. Thus, there is an unblocked path between U and Y not blocked by Z, i.e., $(U \not\perp\!\!\!\perp Y|Z)_G$ which by path indication implies that $(Z \not\perp\!\!\!\perp Y)_{G_{\overline{X}}}$, violating the first IV condition in Definition A2.

Appendix D. Additional Theory and Proofs

Throughout, we let $P(\cdot|\cdot)$ be the conditional probability or density function. As a shorthand, we leave the random variables to be understood from the arguments of P. For example, if $Y(x)$ is discrete, $P[y(x)|u]$ is a shorthand for $\Pr[Y(x) = y(x)|U = u]$.

Auxiliary Lemmas

Lemma 1 (Lemma 4.3 in Dawid (1979))**.** *Let A, B, D, Q be four random variables. If $A \perp\!\!\!\perp B|D, Q$ and $B \perp\!\!\!\perp Q|D$ then $A \perp\!\!\!\perp B|D$.*

Proof. Because $A \perp\!\!\!\perp B|D, Q$, it follows that for all a, b, d, q, we have that

$$\begin{aligned} P(a, b|d, q) &= P(a|d, q)P(b|d, q) \\ &= P(a|d, q)P(b|d), \end{aligned} \tag{A1}$$

where the last line follows from $B \perp\!\!\!\perp Q|D$. Now,

$$P(a, b|d) = \int P(a, b|d, q)P(q|d)dq = \left[\int P(a|d, q)P(q|d)dq\right] P(b|d) = P(a|d)P(b|d),$$

where the second equality is by (A1). □

This lemma is also known as the contraction axiom of conditional independence (Pearl and Paz, 1986). The following lemma is a direct result of Lemma 1.

Lemma 2. *Let A, B, D, Q be four random variables. If $A \perp\!\!\!\perp B|D, Q$ and $A \not\perp\!\!\!\perp B|D$ then $A \not\perp\!\!\!\perp Q|D$ and $B \not\perp\!\!\!\perp Q|D$.*

Proof. Assume by way of contradiction that $B \perp\!\!\!\perp Q|D$. Therefore, by Lemma 1, because $A \perp\!\!\!\perp B|D, Q$ it follows that $A \perp\!\!\!\perp B|D$, which contradicts the assumption. A similar contradiction is received by assuming $A \perp\!\!\!\perp Q|D$. □

1. Negative Controls for IV Designs Under General Definitions

This section presents the proofs of the theoretical results from Section II. We prove versions of the results that are more general in three different ways. First, we discuss IV designs that include control variables. Second, we provide more general definitions for APO and API variables that accommodate multiple threats of which Definitions 1 and 2 are special cases. Third, we provide more general definitions of NCO and NCI (under weaker NCO and NCI assumptions, respectively).

We start by presenting the outcome independence and exclusion restriction assumptions when controls are included.

Assumption A1 (Outcome independence)**.** *$Z \perp\!\!\!\perp Y(z, x)|C$ for all possible z, x values.*

Assumption A2 (Exclusion restriction)**.** $\Pr(Y(z, x) = Y(z', x) = Y(x)|C = c) = 1$ *for all possible z, z', x, c values.*

Similar to the case without controls, outcome independence and exclusion restriction together yield $Z \perp\!\!\!\perp Y(x) | C$.

Alternative Path Variables with Multiple Threats and Controls

In some applications, multiple potential alternative paths can exist between the IV and the outcome. Appendix E.6 presents an example of two distinct variables that affect the outcome and could potentially affect the IV as well and thus may violate outcome independence. To accommodate the possibility of multiple violations of the IV assumptions, we extend Definition 1 and 2. We introduce a random variable V, which represents other potential threats in addition to the threat posed by the alternative path variable U. We also include control variables C to accommodate cases where the IV design is assumed to be valid only when controls are included.

The more general definition for APO variables reads as follows.

Definition A7 (Alternative path outcome variable with multiple threats and controls)**.** *A random variable U is an APO variable conditional on a set of controls C if there exists a random variable V such that the following conditions hold.*

1) *Latent IV validity.* $Z \perp\!\!\!\perp Y(x) | C, U, V$.
2) *Path indication. If* $Z \perp\!\!\!\perp Y(x) | C, V$ *then* $Z \perp\!\!\!\perp U | C, V$.
3) *Direct IV link. If* $Z \perp\!\!\!\perp U | C, V$ *then* $Z \perp\!\!\!\perp U | C$.
4) *V-validity. If* $Z \perp\!\!\!\perp Y(x) | C$ *then* $Z \perp\!\!\!\perp Y(x) | C, V$.

Under this definition, latent IV validity states that the IV is valid conditionally not only on the APO variable U, but also on the additional threat(s) V, and the controls C.

In contrast to Definition 1, this more general version of latent IV validity also holds for U even if V is the actual threat to the identification and U is only an imperfect proxy for it. Therefore, to maintain the same interpretation of an APO variable as posing a threat to IV validity, we replace the condition of path indication from Definition 1 and include two additional conditions. Together, Conditions 2–4 of Definition A7 yield Condition 2 of Definition 1 (conditional on the controls C). However, the three separate conditions ensure that the APO variable is part of the threat itself and not a proxy. Specifically, path indication and direct IV link each rule out a different type of proxy; see Appendix E.7 and Appendix E.8 for counterexamples. The final property, V-validity, is a more technical requirement for the variable V that ensures V represents other threats. It states that an IV design that satisfies outcome independence and exclusion restriction conditional on the controls remains valid conditional on V. See Appendix E.9 for a counterexample. If there are no additional threats other than through U, and no control variables, Definition A7 is equivalent to Definition 1.

We similarly extend Definition 2 to allow for additional threats and to include controls C.

Definition A8 (Alternative path instrument variable with multiple threats and controls)**.** *A random variable U is an API variable conditional on a set of controls C if there exists a random variable V such that the following conditions hold.*

1) Latent IV validity. $Z \perp\!\!\!\perp Y(x)|C, U, V$.

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

3) Direct outcome link. If $U \perp\!\!\!\perp Y|Z, C, V$ *then* $U \perp\!\!\!\perp Y|Z, C$.

4) V-validity. If $Z \perp\!\!\!\perp Y(x)|C$ *then* $Z \perp\!\!\!\perp Y(x)|C, V$.

Condition 1 is the same as in Definition A7. Conditions 2–4 together imply a version of Condition 2 from Definition 2 that includes controls. Similar to the theory of APO variables, the condition is decomposed into three independent conditions to exclude proxies that are not themselves part of an alternative path and maintain that V is indeed a part of such a threat (similar to Definition A7).

General Definition of Negative Control Variables

We adapt the definition of NCOs as follows.

Definition A9 (Negative control outcome with controls)**.** *A random variable NC is an NCO if there exists an APO variable U such that the following two conditions hold.*

1) The NCO assumption. If $Z \perp\!\!\!\perp U|C$ *then* $NC \perp\!\!\!\perp Z|\ U, C$.

2) U-comparability. $NC \not\perp\!\!\!\perp U|C$.

Even without controls ($C = \emptyset$), this definition is more general than Definition 3, as it allows for NCOs that were previously excluded. To see this, note that in the case without controls, if $Z \not\perp\!\!\!\perp U$ (and so the design is invalid), the NCO assumption in Definition A9 allows for an association $NC \not\perp\!\!\!\perp Z|\ U$. In this case, the NCO would still be informative about the validity of the IV design since the association between the NCO and the IV, given the APO variable exists only if the design is invalid. Appendix E.10 provides an example of such an NCO. Every variable that satisfies the NCO assumption in Definition 3 trivially satisfies this less restrictive definition.

Next, we generalize the definition of an NCI to include controls and allow for direct associations with the outcome if the IV design is invalid.

Definition A10 (Negative control instrument with controls)**.** *A random variable NC is an NCI if there exists an API variable U such that the following two conditions hold.*

1) The NC assumption. If $U \perp\!\!\!\perp Y|Z, C$ *then* $NC \perp\!\!\!\perp Y|Z, C, U$.

2) U-comparability. $NC \not\perp\!\!\!\perp U|Z, C$.

We are now ready to state the more general version of Theorem 1 and present its proof. This theorem also covers the case without controls by letting C be degenerate.

Theorem A2. *Assume that a random variable NC is an NCO with respect to controls C (Definition A9). If $NC \not\perp\!\!\!\perp Z|C$, then either outcome independence or exclusion restriction is violated. That is, the IV design is invalid.*

Proof. We begin by showing that $NC \not\perp\!\!\!\perp Z|C$ implies that $Z \not\perp\!\!\!\perp U|C$. Else, if $Z \perp\!\!\!\perp U|C$ then by the NCO assumption (see Definition A9) $NC \perp\!\!\!\perp Z|\ U, C$. Based on Lemma 2, $NC \perp\!\!\!\perp Z|\ U, C$ and $NC \not\perp\!\!\!\perp Z|C$ imply that $Z \not\perp\!\!\!\perp U|C$, a contradiction.

Next, from direct IV link if follows that $Z \not\perp\!\!\!\perp U|C$ implies $Z \not\perp\!\!\!\perp U|V, C$. Then, by path indication, we get that $Z \not\perp\!\!\!\perp Y(x)|V, C$. Finally, by V-validity, we have that $Z \not\perp\!\!\!\perp Y(x)|C$.

However, outcome independence (Assumption A1) and exclusion restriction (Assumption A2) together imply that $Z \not\perp\!\!\!\perp Y(x)|C$. Therefore, one of these assumptions must be violated. □

Next, we state the more general version of Theorem 2 and present its proof. This theorem also covers the case without controls by letting C be degenerate.

Theorem A3. *Assume that a random variable NC is an NCI with respect to controls C (Definition A10). If $NC \not\perp\!\!\!\perp Y|Z, C$, then either outcome independence or exclusion restriction is violated. That is, the IV design is invalid.*

Proof. We divide the proof into two cases with respect to the API variable U for which the NCI assumption holds for NC.

First, assume that $U \not\perp\!\!\!\perp Y|Z, C$. In this case, from direct outcome link it follows that $U \not\perp\!\!\!\perp Y|Z, C, V$, and therefore by path indication, $Z \not\perp\!\!\!\perp Y(x)|C, V$. Therefore, by V-validity, we have that $Z \not\perp\!\!\!\perp Y(x)|C$. Hence, either outcome independence or the exclusion restriction do not hold.

We now turn to the other case, where $U \perp\!\!\!\perp Y|Z, C$. By the NCI assumption (Definition A10), we have that $NC \perp\!\!\!\perp Y|Z, C, U$, and by the condition of the theorem, we have that $NC \not\perp\!\!\!\perp Y|Z, C$. Therefore, by Lemma 2 (with $D = \{Z, C\}, Q = U$), we have that $U \not\perp\!\!\!\perp Y|Z, C$, which contradicts the assumption $U \perp\!\!\!\perp Y|Z, C$.

□

We now turn to state and prove a version of Theorem 3, conditional on controls C.

Theorem A4. *Assume that a random variable NC is an NCI with respect to controls C (Definition A10). If, in addition, $NC \perp\!\!\!\perp Z|C$, then if $NC \not\perp\!\!\!\perp Y|C$, then $Z \not\perp\!\!\!\perp Y(x)|C$.*

Proof. Assume by way of contradiction that $Z \perp\!\!\!\perp Y(x)|C$ holds. Because NC is an NCI, it follows from Theorem A3 that $NC \perp\!\!\!\perp Y|Z, C$. Additionally, based on the assumption that $NC \perp\!\!\!\perp Z|C$, Lemma 1 implies that $NC \perp\!\!\!\perp Y|C$, which contradicts the premise. □

Control Variables and Functional Form

Proof of Corollary 1

Proof. The minimized expression can be written as

$$\mathbb{E}[Z - b_C'C - b_{NC}NC]^2 = \mathbb{E}\big[Z - \mathbb{E}[Z|C, NC]\big]^2 + \mathbb{E}\big[\mathbb{E}[Z|C, NC] - b_C'C - b_{NC}NC\big]^2$$

plus a term equaling zero because $\mathbb{E}[Z - \mathbb{E}[Z|C, NC]]$ is zero by the law of total expectation. Since $\mathbb{E}[Z - \mathbb{E}[Z|C, NC]]^2$ does not depend on b_C, b_{NC}, we can write

$$\gamma = \underset{b_c, b_{NC}}{\arg\min}\, \mathbb{E}\big[\mathbb{E}[Z|C, NC] - b_C'C - b_{NC}NC\big]^2.$$

If outcome independence, exclusion restriction, and rich covariates are assumed, (4) holds and $\mathbb{E}[Z|C, NC] = \gamma_C'C$. Hence, we can further write

$$\gamma = \underset{b_c, b_{NC}}{\arg\min}\, \mathbb{E}[\gamma_C'C - b_C'C - b_{NC}NC]^2.$$

The values that minimize this nonnegative expression are $b_C' = \gamma_C'$ and $b_{NC} = 0$ and so the OLS population-level coefficient is $\gamma' = (\gamma_C', 0)$. If $\gamma_{NC} \neq 0$, it must be that (4) does not hold. Therefore, either outcome independence, exclusion restriction, or rich covariates is violated.

□

Proof of Corollary 2

Proof. Let $\widetilde{Z} = Z - C'\mathbb{E}[CC']^{-1}\mathbb{E}[CZ]$ and $\widetilde{NC} = NC - C'\mathbb{E}[CC']^{-1}\mathbb{E}[CNC]$ be the residuals from the linear regressions of Z and NC on C, respectively. By the Frisch–Waugh–Lovell theorem, we can write β_Z as

$$\beta_Z = \frac{COV(\widetilde{Z}, \widetilde{NC})}{Var(\widetilde{Z})}.$$

If $\beta_Z \neq 0$ then it must be that $COV(\widetilde{Z}, \widetilde{NC}) \neq 0$. Define (γ_C', γ_{NC}) to be the population-level solution of the reverse OLS, with Z as the dependent variable (as in Corollary 1). Again, by the Frisch–Waugh–Lovell theorem, we can write γ_{NC} as

$$\gamma_{NC} = \frac{COV(\widetilde{Z}, \widetilde{NC})}{Var(\widetilde{NC})}.$$

Since $COV(\widetilde{Z}, \widetilde{NC}) \neq 0$ it follows that $\gamma_{NC} \neq 0$ as well. By Corollary 1, we have that either outcome independence, exclusion restriction, or rich covariates does not hold.

□

Proof of Corollary 3

Proof. Similar to the proof of Corollary 1, we write the equivalent minimization problem as

$$\theta = \underset{b_Z, b_c, b_{NC}}{\arg\min}\, \mathbb{E}\big[\mathbb{E}[Y|Z, C, NC] - b_Z Z - b_C' C - b_{NC} NC\big]^2.$$

If outcome independence, exclusion restriction, and CSRF are assumed, (5) holds and

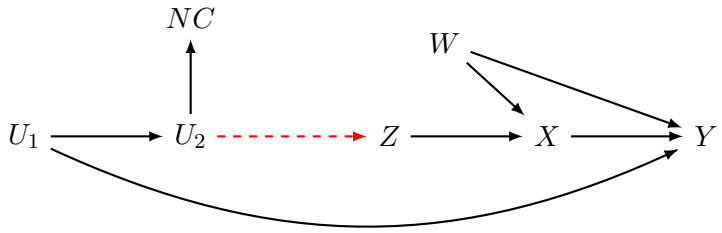
$$\theta = \underset{b_Z, b_c, b_{NC}}{\arg\min}\, \mathbb{E}[\theta_Z Z + \theta_C' C - b_Z Z - b_C' C - b_{NC} NC]^2.$$

The values that minimize this expression are $b_Z = \theta_Z, b_C' = \theta_C'$ and $b_{NC} = 0$ and so the OLS population-level coefficient is $\theta' = (\theta_Z, \theta_C', 0)$. If $\theta_{NC} \neq 0$, it must be that (5) does not hold. Therefore either outcome independence, exclusion restriction, or CSRF is violated.

□

Appendix E. Examples and Counterexamples

1. *Non-Causal APO Variable*



Appendix Figure E1. An illustration of causal and non-causal APO variables

In the example presented in Appendix Figure E1, U_2 is a valid APO variable satisfying path indication without having a direct causal effect on Y (beyond possible effect through the IV). Consider a scenario where Z represents supposedly quasi-random teacher assignment, X is the value added of the actual teacher, and Y is student test scores. The variable W represents some unobserved student characteristic (e.g., their parents' involvement), which affects test score and

also correlates with teacher assignments, as some students switch classrooms after the original assignment.

In this example, U_1 denotes unobserved student ability that directly affects test scores, while U_2 represents detailed unobserved previous exam scores. The dashed arrow indicates the possibility that principals may assign students to teachers based on these detailed scores (e.g., students with low math scores are assigned to specific teachers). The detailed past exam scores satisfy path indication despite not directly affecting future test scores. A path exists between previous detailed scores (U_2) and current scores (Y) because both are affected by ability (U_1).

The variable NC represents aggregated previous test scores (e.g., average past scores in math together with other subjects). In this setting, NC is an NCO, with U_2 as an APO variable. An association between the IV and the aggregated lagged test scores would indicate an alternative path from the IV to the outcome. The presence of this path violates outcome independence, as students with different abilities would sort into different teachers based on previous math scores.

Note that while U_1 is also an APO variable, NC is a valid NCO only with respect to U_2, not with respect to U_1 alone. This is because, conditional on U_1, there is still a correlation between the NCO and the IV ($NC \not\perp\!\!\!\perp Z|U_1$). That is, teacher assignment is not conditionally independent of aggregated test scores.

2. Heterogeneity-Based Violation of Path Indication

(A) Females (B) Males

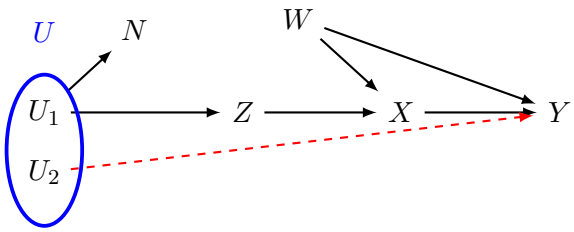
Appendix Figure E2. Violation of path indication (Definition 1) due to heterogeneity

Appendix Figure E2 describes a random variable U associated with both the IV and the potential outcome $Y(x)$. However, U does not qualify as an APO variable because it does not satisfy path indication. The IV affects U for males but not for females, while U affects Y for females but not for males. Path indication is not satisfied because even though the IV satisfied outcome independence and exclusion restriction assumptions ($Z \perp\!\!\!\perp Y(x)$), it remains associated with U ($Z \not\perp\!\!\!\perp U$).

The random variable N is a proxy for U. With sufficient sample size, we would detect $N \not\perp\!\!\!\perp Z$ (due to the effect among males), but this test result does not indicate that the IV is invalid since U is

not an APO variable.

3. Violation of Path Indication: Multivariate Variable



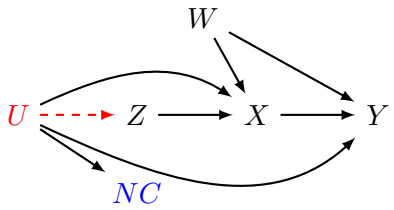
Appendix Figure E3. Violation of path indication (Definition 1) when U has multiple components

Appendix Figure E3 presents an example where $U = (U_1, U_2)$ is a bivariate vector of independent variables. Assume Z represents teacher assignment claimed to be quasi-random, X is the actual teacher's value added, and Y is test scores. The variable W represents some unobserved student characteristic (e.g., their parents' involvement), which also correlates with teacher assignments, as some students switch classrooms after the original assignment. Let U_1 represent having basketball as a hobby and further assume that it is correlated with the IV; for example, one teacher also coaches basketball, so students who list basketball as a hobby are more likely to be assigned to her. However, as seen in Appendix Figure E3, a basketball hobby is independent of test scores ($U_1 \perp\!\!\!\perp Y(x)$). Let U_2 represent having math as a hobby. Students reporting math as a hobby tend to perform better in exams and are randomly distributed between teachers ($U_2 \perp\!\!\!\perp Z$). The basketball and math hobbies are independent ($U_1 \perp\!\!\!\perp U_2$). Finally, assume that N is participation in an extracurricular basketball program, serving as a proxy for U (specifically for U_1).

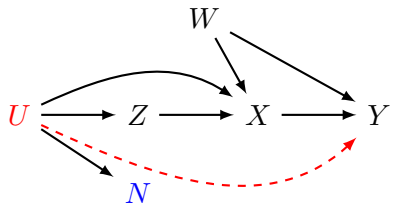
Although the vector U is associated with both the IV and the outcome, it does not qualify as an APO variable because it does not satisfy path indication. The IV satisfies outcome independence and the exclusion restriction assumptions ($Z \perp\!\!\!\perp Y(x)$) despite being correlated with the list of hobbies ($Z \not\perp\!\!\!\perp U$). Therefore, N is not a proper NCO. Even though $N \perp\!\!\!\perp Z|U$, it is still not an NCO because U is not an APO variable.

4. Potential Alternative Path Variables and Association with the Treatment

Path indication for API variables (Definition 2) implies that conditional on the IV (Z), an API variable (U) cannot be associated with the treatment (X). By contrast, there is no such requirement for an APO variable. Appendix Figure E4 illustrates these points. In Panel A, U is a valid APO variable, and the arrow $U \to X$ is allowed: both latent IV validity and path indication hold. Therefore, NC is a valid NCO, and an association between NC and Z implies that the dashed arrow between U and Z exists and indicates that the IV is invalid. Conversely, in Panel B, U is not a valid



(A) $NC \not\perp\!\!\!\perp Z$ implies violation of IV validity. U is an APO variable. NC is an NCI.



(B) $N \not\perp\!\!\!\perp Y|Z$ does not necessarily implies violation of IV validity. U is not an API variable. N is not an NCO.

Appendix Figure E4. Direct effect of the potential alternative path variables on the treatment

API variable. Path indication is violated because $U \not\perp\!\!\!\perp Y|Z$ does not imply $Z \perp\!\!\!\perp Y(x)$. Therefore, N is not an NCI, and $N \not\perp\!\!\!\perp Y|Z$ does not necessarily imply that the IV is invalid. Intuitively, $N \not\perp\!\!\!\perp Y|Z$, even if the IV design is valid, due to the association between U and Y through X. Conditioning on X would not solve this problem because X is a collider. Therefore, $N \not\perp\!\!\!\perp Y|Z, X$ due to the path $N \leftarrow U \rightarrow X \leftarrow W \rightarrow Y$ (Pearl, 2009).

5. Counterexample: A Vector of NCOs That is Not an NCO

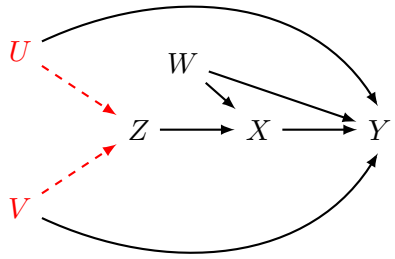
Let R_1, R_2 be two independent Bernoulli random variables with probabilities $\Pr(R_j = 1) = p_j$ and let $p_1 = p_2 = 0.5$. Let U be another Bernoulli random variable, independent of (R_1, R_2). Let Z be the IV, and assume that

$$Z = (R_1 \oplus R_2) + \theta U + \epsilon_Z,$$

where $\oplus$ is the XOR operator. Assume that $Y(x) = x + U + \epsilon_Y$, such that U is an APO variable. The IV design is valid if $\theta = 0$.

Now, assume that there are two observed negative controls $NC_i = U \oplus R_i$ for $i = 1, 2$. Both NC_1 and NC_2 are valid negative controls as they satisfy the assumption $NC_i \perp\!\!\!\perp Z|U$. This is because for $i = 1, 2$, $R_i \perp\!\!\!\perp (R_1 \oplus R_2)$, and therefore $Z \perp\!\!\!\perp R_i|U$. However $(NC_1, NC_2) \not\perp\!\!\!\perp Z|U$ because, conditional on U, Z is associated with $NC_1 \oplus NC_2 = R_1 \oplus R_2$. Therefore, (NC_1, NC_2) does not satisfy the NCO assumption. Indeed, even if the IV is valid, we could still have $Z \not\perp\!\!\!\perp (NC_1, NC_2)$.

A small change in the data-generating process will break some of the independencies discussed above. For example, changing the value of p_1 to something different from 0.5 would imply that $R_2 \not\perp\!\!\!\perp (R_1 \oplus R_2)$. In that case, $NC_2 \not\perp\!\!\!\perp Z$ and NC_2 would no longer satisfy the NCO assumption.



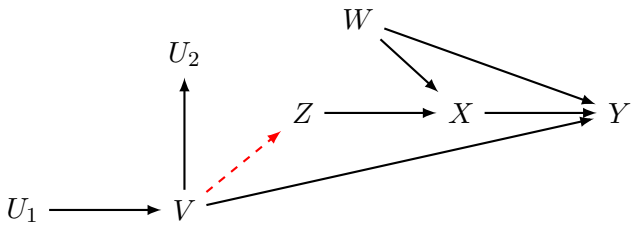
Appendix Figure E6. Multiple threats

6. Multiple Threats

Appendix Figure E6 presents an example of the presence of multiple threats to the validity of the IV design. In this case, the variable U is an APO variable by Definition A7 (taking the control variables C to be an empty set). In this figure, V is an APO variable as well.

For example, assume that Z is the teacher assignment, which is claimed to be quasi-random, X is the value added of the actual teacher, and Y is test scores. The variable W represents some unobserved student characteristic (e.g., their parents' involvement), which also correlates with teacher assignments, as some students switch classrooms after the original assignment. Assume that U is the student's unobserved ability. Assume also that V is principal quality, which is also unobserved. Both U and V might affect the teacher allocation Z, which would generate an alternative path between the IV and the outcome.

7. Direct IV Link Rules out Proxies of V

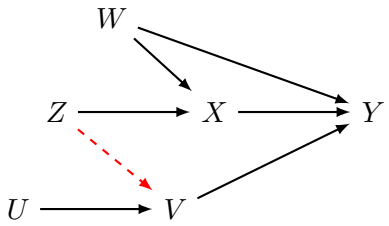


Appendix Figure E7. Violation of direct IV link (Definition A7)

Appendix Figure E7 presents the potential violation of outcome independence through V, as well as two proxies for V, U_1, and U_2. Note that latent IV validity, as stated in Definition A7, holds for either variable (U_1 or U_2) together with V, as the further conditioning on U_1 or U_2 does not invalidate the IV, conditional on V. Note also that for both variables, path indication holds because

if $Z \perp\!\!\!\perp Y(x)|V$ then $Z \perp\!\!\!\perp U_1|V$ (or $Z \perp\!\!\!\perp U_2|V$). However, condition 3 of Definition A7, direct IV link, does not hold. If the IV design is invalid (the dashed line exists), then $Z \not\perp\!\!\!\perp U_1$ while $Z \perp\!\!\!\perp U_1|V$ (and similarly for U_2). Intuitively, we rule out U_1 and U_2 as APO variables because they are only proxies for the variable creating the threat to IV validity.

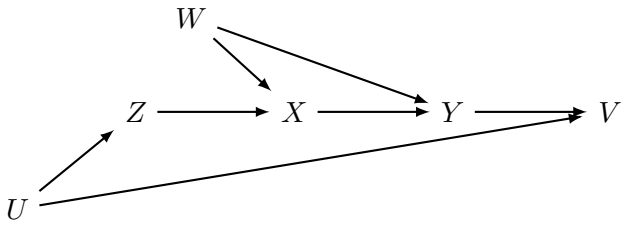
8. Path Indication Rules Out Proxies of V



Appendix Figure E8. Violation of path indication (Definition A7)

Appendix Figure E8 presents a potential violation of exclusion restriction through the variable V, as well as a proxy for V, labeled as U. While V itself is an APO, its proxy U is not. Note that latent IV validity holds for U, V jointly, as the further conditioning on U does not invalidate the IV design once we have conditioned on V. Note also that direct IV link holds because $Z \perp\!\!\!\perp U$. However, U is not an APO variable because it does not represent a threat to IV validity. Condition 2 of Definition A7, namely, path indication, does not hold. Specifically, if the IV design is invalid (the dashed line exists), $Z \not\perp\!\!\!\perp U|V$ while $Z \perp\!\!\!\perp Y(x)|V$. In the language of DAG terminology, V is a *collider* (Pearl, 2009), and conditioning on it creates a dependence between U and Z.

9. Violation of V-validity



Appendix Figure E9. Violation of V-validity

Appendix Figure E9 presents a situation with no valid APO variable. We examine U as a candidate APO variable and consider V in the DAG as the potential V in Definition A7. We see that latent IV validity holds: while $Z \not\perp\!\!\!\perp Y(x)|V$, because V is a common effect (a collider) of U and Y, controlling

for U in addition to V blocks the flow of association (Pearl, 2009), resulting in $Z \perp\!\!\!\perp Y(x)|U,V$. Path indication holds because $Z \not\perp\!\!\!\perp Y(x)|V$. Direct IV link also holds because of the effect of U on Z. However, it is clear U should not be an APO variable. An association between Z and U does not imply that the IV design is invalid. This is where $V-$validity comes to the rescue. The IV satisfies $Z \perp\!\!\!\perp Y(x)$, but, as previously noted $Z \not\perp\!\!\!\perp Y(x)|V$, due to V being a common effect of both variables. In this case, no other alternative to V exists to satisfy Definition A7. Therefore, U is not an APO variable.

10. NCO Potentially Affecting the IV

(A) $Z \perp\!\!\!\perp U$ (B) $Z \not\perp\!\!\!\perp U$

Appendix Figure E10. Conditional dependence between an NCO and the IV

Appendix Figure E10 presents a scenario in which if the IV is invalid, it could also be associated with the NCO, not through the APO variable. For concreteness, consider the case of studying the effect of teacher quality (X) on test scores (Y). The IV (Z) is claimed to be a random assignment of teachers. Unobserved ability (U) is the APO variable. In the case of random assignment, ability has no association with the IV (Panel A). However, there is a concern that the initial assignment was not random in practice. In Panel B, random assignment did not take place, and so other considerations could have impacted the IV, including unobserved ability U. Moreover, it is possible that proxies for unobserved ability, such as lagged test scores (NC), were used directly in the assignment process as well. In this case, $NC \not\perp\!\!\!\perp Z|U$, and so the lagged outcome does not satisfy Definition 3. However, if the IV design is valid, and $Z \perp\!\!\!\perp U$, the condition $NC \perp\!\!\!\perp Z|U$ is satisfied. Hence, NC is an NCO based on the broader Definition A9, defined in Appendix D.1. Indeed, in this case, if $NC \not\perp\!\!\!\perp Z$, the design is invalid ($Z \not\perp\!\!\!\perp Y(x)$).

Appendix F. Details of Implementation of Negative Control Tests Using Data from Prior Studies

This section provides additional details for the analysis from Section III, which implements our proposed methods on IV designs used in prior studies. Appendix Table A1 summarizes information

about the key variables in each study. We are grateful to the authors of these prior studies for publicly posting their data and code. In each case, we first used the publicly posted data to replicate the related original study's results (this step is not further discussed here). We then applied our additional negative control falsification tests. Replication code and data are included in the supplementary materials.

1. Implementation Details for Autor, Dorn and Hanson (2013)

Sample Construction. For this analysis, we use the original study's data from Autor, Dorn and Hanson (2013, henceforth ADH), which is taken from the US Census. The unit of analysis is a commuting zone. The sample included 722 commuting zones.

Main Variables. For each commuting zone, we observe all variables from the original study's replication data and additional variables not used in the original study, some of which we use as NCOs in our current analysis. The treatment and IV are built as shift-share variables, weighting change in Chinese import by industry where weights are the local industry shares in the commuting zone. The treatment uses Chinese imports in the US and the IV uses Chinese imports in other developed countries to avoid endogeneity. We focus on the analysis for the years 2000–2007. The treatment and IV are the shift-share difference in Chinese imports between the years 2007 and 2000. The control variables are the lagged year 2000 values. Note that ADH also used another version of the IV, measured between 1990–2000. We do not evaluate this version because it would not allow us to use the large set of variables from 1990 as NCOs.

Original Falsification Tests. ADH conducted falsification exercises to evaluate the concern that the decline in US manufacturing employment in commuting zones with high exposure to Chinese imports might have occurred for reasons unrelated to Chinese imports. They regress past changes in the manufacturing employment share on future changes in import exposure (See Columns (4)–(6) of Table 2 in ADH). This relationship was found to be significant only for 1970–1980, but not for 1980–1990 or 1970–1990. The significant specification yielded a coefficient with the opposite sign. We replicated this analysis and obtained a similar result. The p-value is reported in Column (1) of our Table 3. This original exercise is similar in spirit to our proposed approach, although it uses the different negative controls separately and not jointly. It also uses a 2SLS specification for estimation, not the reduced form.

The remainder of this section discusses additional falsification tests that we performed using alternative negative control variables sourced from the original replication data.

Additional NCOs. We use 52 NCOs in our falsification analysis. These include the NCOs that were originally used by ADH (lagged changes in manufacturing employment) and all variables measuring labor market conditions in 1990. In particular, we include the share of workers who were employed in manufacturing, employed in non-manufacturing, unemployed, and not in the labor force, separately for males, females, college educated, non-college educated, and for three different

age groups; the share who received SSDI; average log weekly wages in manufacturing and in non-manufacturing; average household total income and average household wage; total population and size of the workforce; levels of transfers per capita for medical benefits, federal income assistance, unemployment benefits, TAA benefits, education/training assistance, SSA retirement benefits, SSA disability benefits, other assistance, and total individual transfers.

Implementation Details. We use the same sampling weights used by ADH in the original study (`timepwt48`). We also follow ADH and cluster standard errors by states (`statefip`).

In Column (1) of Table 3, we use a single NCO that was used in the original analysis, namely the change in manufacturing employment between 1970–1980. We replicated the ADH analysis, which regressed past outcomes (1970) on future treatments (years 1990 and 2000 averaged), instrumented by future IVs (see Column (4) of Table 2 in ADH). We report the p-value of the coefficient on the treatment. In Column (2), we perform a similar analysis by regressing the 1970 outcome on the year 2000 IV (e.g., reduced form), including the full set of 16 control variables (as in Column (6) of Table 3 in ADH).

2. Implementation Details for Deming (2014)

Sample Construction. We use the original study's data from a public school choice lottery in Charlotte-Mecklenburg, North Carolina. The unit of analysis is the individual student. The sample includes 2,343 students.

Main Variables. We use Deming's VAM estimates from the mixed-effects specification, controlling for past test scores.[27] Based on the replication code, we can write the IV as

$$IV_i = L_i VAM_i^1 + (1 - L_i)\, VAM_i^N \tag{A1}$$

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. These variables are included in the original study's replication data.

Control variables include lagged test scores from the year 2001–2002 as well as lottery fixed effects (i.e., a categorical variable for every choice of school ranking). Following Deming (2014), the test scores include the math and reading test scores in nominal, quadratic, and cubic values, and an indicator of missing values.

NCOs. The original study did not report any falsification tests. We perform falsification analysis using lagged test scores from earlier school years (1998–2001) that were included in the replication data but not included as controls in the study (see the control variables definition) and lagged outcome (`testz2002`; i.e., 2002 test scores). We also used the VAM of the three schools that the

[27] The original study included richer specifications (models 3–4 in the original study) that controlled for individual characteristics, which were not made publicly available due to privacy constraints.

student applied to in the lottery and the neighborhood school's VAM. In total, we used 37 NCOs.

Implementation Details. Following the original paper, all our analyses are unweighted. In the analysis with a single NCO, we replace the outcome with lagged test scores (from 2001–2002) in the reduced form. In the F-test and multiple linear tests with Bonferroni correction, we perform a fixed-effect regression of the IV on the NCOs with the `lottery_FE` variable. In the GAM models, fixed effects are accounted for by taking `lottery_FE` as a categorical variable without a smooth term.

Additional Analysis. In Appendix Figure A2, we show the correlation of each NCO with the outcome and the IV. Before calculating each correlation, we residualized the NCO and the IV or the outcome by the control variables.

In an unreported analysis, we replicated the main 2SLS results using L_i, the raw lottery outcome, as an alternative IV. The point estimates remained statistically unchanged, although standard errors were larger.

3. Implementation Details for Nunn and Qian (2014)

Sample Construction. We use the study data, which consists of annual panel data of 125 non-OECD countries over 36 years. The sample includes 4,572 observations.

Main Variables. The IV of the study is the US wheat production from the previous year. We limit our analysis to the main outcome variable of the study, which is the intrastate conflict indicator. We utilize the extended set of 238 control variables (as in the "baseline specification" in Table 2 of Nunn and Qian (2014)).

NCIs. As in the original study, we used a set of ten NCIs. The NCIs are the lagged US production of various products that are not sent as foreign aid.

Original Falsification Tests. Nunn and Qian (2014) performed a falsification test (Table 5 in Nunn and Qian) with the aforementioned NCIs by estimating the reduced form equation

$$Y_i = NCI_i^j + IV_i + C_i + \epsilon_i$$

for each of the ten NCI^j and the "baseline specification" of the control variables.

Implementation Details. In all analyses, we follow Nunn and Qian (2014) and cluster standard errors by country.

Additional Analysis. We can also implement a GAM model with linear controls. This test rejects the null hypothesis. The rejection is driven at least in part by a violation of the unnecessary CSRF Assumption (Assumption 4). To test the functional form, we implement Ramsey's RESET test for misspecification with quadratic and cubic fitted values for the reduced form equation. This test results in a p-value lower than 1%, implying a misspecification. However, the large number of

control variables does not allow for estimating a GAM model with smooth controls as well or for including interactions of the control variables. Therefore, we cannot assess IV validity separately.

4. Implementation Details for Ashraf and Galor (2013)

Sample Construction. The study data consists of a sample of 145 countries.

Main Variables. The outcome of the study is the historical population density, which is defined as the log population density in 1500 CE. The main IV is the migratory distance from Addis Ababa. We use the same set of four control variables included in the study.

NCIs. We use the same three NCIs as in the original study, which are the migratory distances from London, Tokyo, and Mexico City.

Implementation Details. We follow Ashraf and Galor (2013) and include a quadratic polynomial for both the IV and the NCIs.