Negative Control Falsification Tests for Instrumental Variable Designs*

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

May 7, 2024

Abstract

We develop theoretical foundations for widely used falsification tests for instrumental variable (IV) designs. We characterize these tests as conditional independence tests between negative control variables — proxies for potential threats — and either the IV or the outcome. We find that conventional applications of these falsification tests would flag problems in exogenous IV designs, and propose simple solutions to avoid this. We also propose new falsification tests that incorporate new types of negative control variables or alternative statistical tests. Finally, we illustrate that under stronger assumptions, negative control variables can also be used for bias correction.

*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, Princeton University, Stanford University, Technion, and Tel Aviv University 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.

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

The identification assumptions in instrumental variable (IV) designs cannot be tested directly. To assess these assumptions indirectly, researchers often use falsification ("placebo") tests. Surveying the most highly cited papers published in five leading economics journals over the past decade, we find that 51% of the papers using IV have implemented such falsification tests. The large majority of falsification tests fall into two categories. Among the studies that conducted falsification testing, 72% tested that the IV is not associated with certain variables, which we call *negative control outcomes* (NCOs). For example, they tested that the IV is not correlated with the lagged outcome. Similarly, 25% tested 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 resembling the IV but not affecting the treatment. Despite widespread use in IV designs, these negative control tests lack a comprehensive theoretical foundation.[1]

In this paper, we develop a theory of negative control testing for IV designs. We introduce a formal definition for threats to IV exogeneity. Building on this definition, we characterize the conditions that proxy variables for such unobserved threats must meet to qualify as negative controls. The theory concludes that negative control variables can test IV exogeneity using (conditional) independence tests. Our theory highlights two pitfalls prevalent in current practice. First, NCI tests often (but not always) require conditioning on the IV. This crucial step is frequently overlooked in empirical implementations, potentially leading researchers to find false problems in valid IV designs. Second, prevalent negative control tests may flag problems in exogenous IV designs due to violations of functional form assumptions. We propose ways to test IV exogeneity distinctly from functional form assumptions, which are often either unnecessary or replaceable. Further, our theoretical framework suggests novel negative control variables and testing methods underutilized in current empirical work. Finally, we illustrate that when a bias is detected, NCOs can also be used for bias correction, albeit under stricter assumptions.

We distinguish between two categories of negative control tests. The first category, NCO tests, examines whether NCO variables are independent of the IV. Figure 1 provides two examples of such tests.[2] Panel A illustrates a case of a potential violation of the independence assumption (see, e.g., Abadie, 2003). For concreteness, consider the context of Martin and Yurukoglu (2017), who evaluated the impact of Fox News viewership (X) on Republican vote shares (Y), using cable channel positions (Z) as an IV for viewership. A possible concern is that the unobserved level of conservativeness of the local population (U_1) may affect the cable networks' channel positioning strategies. In such a case, there is an alternative path

[1] In epidemiology and bio-medical settings, researchers use similar terminology for falsification tests for potential confounding of an exposure–outcome relationship (Lipsitch et al., 2010; Shi et al., 2020).

[2] Throughout the paper, we use directed acyclic graphs (DAGs) to visualize complex structures, as advocated by Imbens (2020). We do not use DAG theory (e.g., Pearl, 2009) in our theoretical framework.

between the channel position and Republican voting, rendering the design invalid. Because the level of conservativeness is not directly observed, Martin and Yurukoglu indirectly test if it is associated with channel position by using an observed proxy. Specifically, they use local Republican vote share in 1996 as an NCO (denoted by NC_1 in Figure 1). If voting in 1996 is associated with channel position, an alternative path exists between the IV and the outcome, violating the independence assumption. Panel B of Figure 1 illustrates how a different NCO could test the exclusion restriction assumption. Panel A of Table 1 describes further examples of applications of NCO tests in economic research.

The second category of negative control tests, NCI tests, evaluates conditional associations between NCI variables and the outcome. Panel C of Figure 1 provides an example of an NCI test for potential violation of the independence assumption. For concreteness, consider the context of Nunn and Qian (2014), who study the effect of US food aid (X) on conflicts in the recipient country (Y), using US wheat production (Z) as an IV for aid. In this example, the threat to identification could be unobserved weather conditions (U_3). Here, it is known that the threat, weather, affects the IV. The question is whether it also affects the outcome (not through the treatment). Since weather conditions are not fully observed, Nunn and Qian indirectly test for their association with the outcome by using an observed proxy. Specifically, the US production of other crops that are also affected by similar weather conditions but are not used for food aid (e.g., oranges) serves as an NCI (NC_3): if orange production is associated with conflicts, conditional on wheat production, then there exists an alternative path between the IV and the outcome, violating the independence assumption. Panel D of Figure 1 illustrates how NCIs can also help evaluate the exclusion restriction assumption. Panel B of Table 1 lists additional examples of NCI tests in economic research.

We develop a theoretical framework that formalizes these tests. To this end, we introduce the concept of an *alternative path variable*, which represents a potential threat to the identification (such as the U variables in Figure 1). These variables potentially establish a path from the IV to the outcome. If such a path exists (e.g., if the dashed red lines in Figure 1 are present), IV exogeneity does not hold.

Building on this definition, we define a negative control variable as a proxy for an alternative path variable. This proxy is defined such that it can indicate the existence of the alternative path using a (conditional) independence test between the NCO and the IV or between the NCI and the outcome. In particular, the definition of the NCO rules out variables that are uninformative of the design validity because they are associated with the IV not through the alternative path variable (e.g., variables directly affected by the IV). Similarly, the definition of the NCI rules out variables associated with the outcome not through the alternative path variable or the IV (e.g., variables directly affecting the outcome).

Our theory exposes two potential pitfalls that are prevalent in current practice. First, the common implementation of NCI tests in applied economics studies is expected to find false problems in exogenous IV designs (with a large enough sample size). In 94% of the papers that implemented an NCI test, the NCI was a pseudo-IV, i.e., a variable similar to the IV, but that does not affect the treatment (e.g., orange production instead of wheat production). In 92% of these papers, the original IV is substituted with the pseudo-IV (the NCI) in the reduced form equation. The NCI test assesses whether the pseudo-IV is correlated with the outcome. This approach neglects the potential correlation between the pseudo-IV and the original IV. Since the original IV affects the outcome through the treatment, the pseudo-IV might also correlate with the outcome, even if the IV design is exogenous.[3] Using the previous example, depicted in Panel C of Figure 1, orange production (NC_3) and wheat production (Z) are strongly correlated, as both are affected by similar weather conditions (U_3). As a result, the production of oranges would be correlated with conflict levels (Y) even if the IV is exogenous (i.e., the dashed red line does not exist). This is because orange production is correlated with wheat production, and wheat production is affecting conflicts (through its effect on food aid).

This problem can be solved by controlling for the original IV in the NCI test. In the previous example, conditional on wheat production (Z), orange production (NC_3) and conflict levels (Y) would only be correlated if the IV is not exogenous (i.e., the dashed red line in Panel C of Figure 1 exists). Replicating NCI tests from recent economics papers, we show that once we condition on the IV, false problems in the IV designs are no longer detected. If NCI tests are conducted correctly, they would find fewer false problems in exogenous designs, and could potentially be used more frequently.

Our theory also shows that in some NCI settings, conditioning on the IV is unnecessary. Conditioning on the IV is not required if the IV and the NCI are independent. Such independence is possible when the NCI tests a violation of the exclusion restriction. However, it is not satisfied in pseudo-IV settings, which are substantially more prevalent. In these settings, the NCI is intentionally similar to the original IV (as in the examples in Table 1).

The second pitfall is that both NCO and NCI tests could flag problems in exogenous IV designs due to misspecified functional form. In 2SLS specifications, researchers must choose a functional form for the IV and the control variables in the reduced form equation, typically assuming a simple linear-additive model. This structure often carries over into the execution of negative control tests. Consequently, even exogenous IV designs could fail these tests due to violations of functional form assumptions. However, unlike exogeneity, these functional form assumptions are often replaceable and, in some cases, even unnecessary for identifying causal effects.

To address this problem, we propose using alternative negative control tests that do not

[3] A related problem is common in certain negative control tests in epidemiology (Shi et al., 2020).

rely on these functional form assumptions. However, since these tests make fewer parametric assumptions, they typically require more data to achieve desirable power. We use simulations to demonstrate that in contrast to commonly used methods, alternative tests can test IV exogeneity without testing the functional form assumptions.[4]

We also explore ways to broaden the application of negative control tests, drawing from our theory. We identify new types of negative control variables, some of which have yet to be commonly employed in empirical research. For example, variables that cause the IV could serve as valid NCIs (e.g., observed weather conditions can serve as NCIs when the IV is wheat production). We also propose additional diagnostic tools that may assist in detecting the problem in the IV design when the negative control test rejects the null hypothesis.

Finally, we show that under stronger assumptions, NCOs can also be used for bias correction in IV designs. We focus on a scenario where outcome measurements from a period prior to the initiation of the IV and treatment serve as an NCO. In this scenario, the NCO can also be used for bias correction using a *difference-in-Wald* estimator. This estimator replaces the outcome variable with the difference in outcomes between the two periods. This estimator is consistent for the treatment effect under additional assumptions, including treatment effect homogeneity and an IV version of parallel trends. In more rare scenarios, NCOs can be used for bias correction even in the presence of a heterogeneous treatment effect.

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 et al., 2023; Chyn et al., 2024). Previous work has also discussed robustness tests, not specific to IV, based on varying the set of controls (Altonji et al., 2005; Oster, 2019; Diegert et al., 2022). Eggers et al. (2023) discuss the usage of placebo tests in the social sciences more broadly. We contribute to this literature by providing a theoretical foundation for the most common type of falsification test for IV designs.

This paper also contributes to the growing literature on negative controls by extending their theory to IV designs. Davies et al. (2017) use negative control in IV contexts without developing a theoretical framework for such an approach. We find several important differences in the theoretical framework of negative controls in IV settings compared to non-IV settings. In particular, tests for unconditional independence between a negative control and the outcome are unique to IV settings. Furthermore, a rapidly growing literature also discusses how negative control variables can be used for bias correction in non-IV settings (Sofer et al., 2016; Shi et al., 2020; Tchetgen Tchetgen et al., 2020). When extending bias correction to IV designs, we show that it requires stronger assumptions. We also contribute

[4] **R** code implementing these methods and examples for their usage in our simulation is available from `https://github.com/barwein/NC_for_IV`.

to this literature by formally defining alternative path variables, which are central to negative control theory. Moreover, we extend this definition to cases of multiple identification threats.

The rest of this paper proceeds as follows. Section 1 surveys the current practice of falsification tests for IV designs. Section 2 develops the theory underlying negative control tests. Section 3 discusses commonly used as well as underutilized testing procedures. Section 4 provides guidance for practitioners on implementing and interpreting negative control tests. It also demonstrates the key findings of the paper on recent empirical studies and shows that they affect the results. Section 5 discusses bias correction. Section 6 concludes.

1 Survey of Current Practice

In this section, we 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. The results are summarized in Table 2. We highlight six key findings from this survey.

First, falsification tests are widely used in IV analyses. Approximately half (51%) of all surveyed articles employ some form of falsification test (Column (2) of Table 2).

Second, most falsification tests fall within the negative control framework outlined in this paper. We categorize negative control tests into two types: *negative control outcome* (NCO) tests, which examine associations of the IV with variables it should not be associated with (e.g., lagged outcomes). NCO tests were used in 72% of papers with some falsification tests (Column (3)). The second type is *negative control instrument* (NCI) tests, which examine the association of the outcome variables with variables it should not be associated with. Such tests were used in 25% of papers that included some falsification tests (Column (4)). All other types of falsification tests combined were implemented in 21% of the papers (Column (5)). A list of these less common falsification test types is given in Appendix B.

Third, current applied work usually restricts itself to two simple types of negative control tests, which rely on the 2SLS functional form assumptions. Most NCO tests implement pseudo-outcome tests (Athey and Imbens, 2017). These tests involve estimating a revised reduced form equation by using an alternative outcome (e.g., a lagged outcome) and testing if it is unrelated to the IV. Pseudo-outcome tests account for 56% of all NCO tests. The remaining NCO tests often follow a similar logic.[5]

For NCI tests, applied work predominantly uses pseudo-IV tests. These tests replace

[5] For example, balance tables often regress various NCOs on the IV. Most balance tables are not classified as pseudo-outcome tests in our analysis since they do not use the same specification as the reduced form.

the original IV with a similar variable that does not affect the treatment (the pseudo-IV) in the reduced form equation. Out of all the papers with NCI tests surveyed, 94% used a pseudo-IV test.

Fourth, most reported NCI tests are implemented incorrectly. As we later discuss, pseudo-IV tests should always control for the original IV. In practice, only 19% of the papers surveyed reported doing this. With sufficient statistical power, this error leads to finding false problems in valid IV designs. Presumably, more NCI tests were conducted incorrectly, finding false problems in valid designs, and therefore were not reported.

Fifth, 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)), and 35% of the papers used only one negative control variable. These findings suggest that researchers use only a subset of the valid and relevant negative controls available in their data. As we demonstrate in Section 4, our theory can guide a more 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.

Finally, none of the surveyed papers used negative controls for bias correction. However, negative control test failures often prompted authors to discuss the sign and magnitude of the association. This suggests that authors view sign and magnitude as meaningful for evaluating the bias direction and size. In Section 5, we outline additional assumptions that are required for such an interpretation.

2 Theory of Negative Controls in IV Settings

In this section, we present the theory of negative control tests for IV designs. We start with formal definitions of an IV setting and IV exogeneity in particular. Subsequently, we discuss NCO tests and how they can examine exogeneity. Then, we discuss NCI tests and show that they typically require conditioning on the IV. Finally, we discuss the inclusion of control variables and show that both types of tests can depend on functional form assumptions that are modifiable and sometimes unnecessary.

2.1 Setup

Consider i.i.d. units indexed by $i = 1, \dots, n$. 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 set to the values z and x, respectively.[6] We make the standard assumption that the observed outcome Y_i is given by $Y_i = Y_i(Z_i, X_i)$. Because the units are assumed

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

to be i.i.d., we omit the subscript i when it improves clarity. Unless stated otherwise, all variables may be discrete or continuous.

A valid IV design relies on several assumptions. The negative control tests we discuss in this paper focus on a subset of these assumptions. The first assumption, *outcome independence*, maintains that IV assignment is independent of the potential outcomes.

Assumption 1 (Outcome independence)**.** *For all z, 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. The examples in Panels A and C of Figure 1 illustrate a potential violation of outcome independence.

The second assumption, *exclusion restriction*, maintains that the IV has no direct effect on the outcome.

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

The examples in Panels B and D of Figure 1 illustrate potential violations of this assumption.

Together, outcome independence and exclusion restriction imply *IV exogeneity*, that is,

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

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

To identify a causal effect using an IV design, additional assumptions are also necessary. Depending on the specific design, such assumptions may include treatment independence, relevance, and monotonicity. However, the negative control tests presented in this paper do not test these other assumptions.

2.2 Negative Control Outcomes

2.2.1 Alternative Path Outcome Variables

Our theory characterizes negative controls as proxies for threats to the IV design's validity. To formalize the notion of a threat, we introduce the concept of an *alternative path variable*. This is a variable that is part of a suspected alternative path between the IV and the outcome that, should such a path exist, would violate IV exogeneity. We begin by formalizing the first type of identification threat by defining *alternative path outcome* (APO) variables.

Single Violation. For simplicity, we begin by assuming that only one potential threat to IV exogeneity exists. Figure 1 illustrates such a threat to outcome independence (Panel A) and exclusion restriction (Panel B). In these examples, the APO variable is represented by U. We later provide a more general definition of an APO variable, allowing for multiple potential threats, for which the following definition is a special case.

Definition 1 (Alternative path outcome variable with a single violation). *A random variable U is an APO variable if the following two conditions hold.*

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

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

The first condition, latent IV exogeneity, posits that had we observed and conditioned on the APO variable, the IV design would have been valid. Under this condition, imperfect proxies for identification threats cannot be APO variables, as controlling for an imperfect proxy does not make the IV and the potential outcome conditionally independent. Using the previous example of Martin and Yurukoglu (2017), Republican vote share in 1996 is only a proxy for the threat to IV validity (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 exogeneity condition. Latent IV exogeneity is analogous to the latent exchangeability assumption appearing in recent literature on negative controls in epidemiology (Shi et al., 2020) and statistics (Tchetgen Tchetgen et al., 2020).

The second condition, path indication, states that an exogenous 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 C.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 C.2 and Appendix C.3.

Multiple Violations. In some applications, multiple potential alternative paths can exist between the IV and the outcome. Appendix C.4 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 IV exogeneity, we extend Definition 1. We introduce a random variable V, which represents other potential threats in addition to the threat posed by the APO variable U.

Definition 2 (Alternative path outcome variable)**.** *A random variable U is an APO variable if there exists a random variable V such that the following conditions hold.*

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

Under this definition, latent IV exogeneity states that IV exogeneity holds conditional not only on the APO variable U, but also on the additional threat(s) V. In contrast to Definition 1, this more general version of latent IV exogeneity 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 a threat to IV exogeneity, we replace the condition of path indication from Definition 1, and include two additional conditions. Combining Conditions 2–4 yields Condition 2 of Definition 1. However, the three separate conditions ensure that the APO is the threat itself and not a proxy. Specifically, path indication and direct IV link each rule out a different type of proxy; see Appendix C.5 and C.6 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 exogenous IV remains exogenous conditional on V. See Appendix C.7 for a counterexample. If there are no additional threats other than U, Definition 2 is equivalent to Definition 1.

2.2.2 Negative Control Outcome Assumption

Building on the definition of an APO variable, we are now ready to formalize the assumption required for a random variable to serve as a *negative control outcome*.

Definition 3 (Negative control outcome)**.** *A random variable NC is an NCO if it satisfies the NCO assumption: There exists an APO variable U such that*

$$Z \perp\!\!\!\perp NC|U.$$

This assumption guarantees that any association between the IV and the NCO is through the APO variable U. Panels A and B of Figure 1 depict two examples of NCOs that satisfy this assumption. The NCO assumption is violated for variables that are directly related to the IV, not through an APO variable. Appendix C.8 shows two examples of such violations.

The NCO assumption formalizes ad hoc discussions about the validity of NCOs. For example, Guidetti et al. (2021) investigate the use of non-respiratory hospital admissions as NCOs in IV studies on the impact of air pollution exposure. One might expect that these admissions would only correlate with flawed IVs for air pollution. However, Guidetti et al. demonstrate otherwise. They find that air pollution indirectly increases non-respiratory admissions through hospital congestion caused by a surge in respiratory admissions. Therefore, non-respiratory admissions are not informative about IV exogeneity 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 admission rates violate the NCO assumption.

The NCO assumption can be weakened to include more variables that are informative about the exogeneity of the IV design. In Appendix D we offer a more general definition of NCO that allows for direction associations with the IV, not through the APO variable, if the design is not exogenous.

2.2.3 Negative Control Outcome Test

A *negative control outcome test* (NCO test, for short) is any statistical test of independence between the IV and an NCO. The null hypothesis is $H_0 : Z \perp\!\!\!\perp NC$. Statistical testing procedures are reviewed and compared in Section 3. The following theorem states that rejecting this null implies a violation of IV exogeneity.

Theorem 1. *Assume that a random variable NC satisfies the NCO assumption. If $Z \not\perp\!\!\!\perp NC$, then either outcome independence or exclusion restriction is violated. That is, the IV design is not exogenous.*

The proof is given in Appendix D.1.1 for the more general case, discussed in Section 2.4, which also includes control variables in the design. For the case without controls and a single violation, 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., IV exogeneity does not hold.

Although a failed NCO test indicates that the IV design is not exogenous, the converse is not always true. An IV design that is not exogenous might still pass an NCO test. This would happen if there are other alternative paths from the IV to the outcome that are not captured by the NCO, or due to lack of statistical power. The same logic would apply to NCI tests, which we now turn to discuss.

2.3 Negative Control Instruments

2.3.1 Alternative Path Instrument Variables

As with NCOs, NCIs are typically proxies for threats to identification. We define such threats as *alternative path instrument* (API) variables. Much like an APO variable, an API variable is also part of a suspected alternative path between the IV and the outcome that could violate IV exogeneity. However, API variables are known to be associated with the IV and researchers are concerned about 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 researchers are concerned about their possible association with the IV.

Figure 1 demonstrates two such cases. In Panel C, the IV is known to be non-random, as it is affected by an unobserved API variable U_3. In Panel D, the IV is known to affect another API variable, U_4, in addition to its effect on the treatment. In these examples, IV exogeneity depends on whether these API variables directly affect the outcome Y.

Formally, in the case of a single violation, an API variable satisfies the following definition.

Definition 4 (Alternative path instrument variable with a single violation)**.** *A random variable U is an API variable if the following two conditions hold.*

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

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 exogeneity, is exactly as before. The difference between API and APO variables is encapsulated in the path indication condition. For API variables, this condition requires that if IV exogeneity is satisfied ($Z \perp\!\!\!\perp Y(x)$), then the API variable must be independent of the observed outcome conditional on the IV ($U \perp\!\!\!\perp Y|Z$). This condition implies that an association between the API variable and the outcome, not through the IV, indicates that the IV is not exogenous. Typically, path indication is satisfied when the API variable is associated with the IV. Therefore, if the API variable is also directly associated with the outcome, an alternative path between the IV and the outcome exists.

Path indication also 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 when IV exogeneity is satisfied. This is different from APO variables that could be associated with the treatment (even conditional on the IV). See Appendix C.9 for an example and further discussion of this issue.

The definition of an API variable with a single violation (Definition 4) can be extended to settings where additional violations are potentially present. As with APO variables, this extension requires more nuanced assumptions. The API definition also generalizes to an IV design that includes control variables. Both extensions are presented in Appendix D.1.2.

2.3.2 Negative Control Instrument Assumption

A *negative control instrument* is a variable satisfying the following NCI assumption.

Definition 5 (Negative control instrument)**.** *A random variable NC is an NCI if it satisfies the NCI assumption: There exists an API variable U such that*

$$Y \perp\!\!\!\perp NC|Z, U.$$

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. These two differences reflect that NCI tests, defined below, test for an association with the outcome, and not with the IV. The third difference is that the independence is conditional on the IV as well. This is because in exogenous IV designs the NCI can be associated with the outcome through the IV, as we discuss in the next section.

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

As with NCOs, in Appendix D.1.2 we provide a more general definition of NCI that allows for direct associations with the outcome if the design is not exogenous.

2.3.3 Negative Control Instrument Test

A *negative control instrument test* examines whether the outcome and the NCI are independent, conditional on the IV. Formally, the statistical test is for the null hypothesis $H_0 : Y \perp\!\!\!\perp NC|Z$. 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 satisfies the NCI assumption. If $Y \not\perp\!\!\!\perp NC|Z$, then either outcome independence or exclusion restriction is violated. That is, the IV design is not exogenous.*

The proof is given in Appendix D.1.2 for the more general case that includes control variables and multiple threats.

Conditioning on the IV is typically required, 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 the IV design is valid. In the previously discussed example of Nunn and Qian (2014), the production of oranges (the NCI) is associated with conflicts (the outcome), as both are associated with the production of wheat (the IV).

However, in some cases, conditioning on the IV is not required. When $Z \perp\!\!\!\perp NC$, researcher can use an unconditional independence test for the null $H_0 : Y \perp\!\!\!\perp NC$, as formalized in the next theorem.

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

The proof is given in Appendix D.1.2.

Unconditional NCI tests may be valid mostly when considering violations of the exclusion restriction assumption. Panel A of Figure 2 shows an example of such a case. The IV (Z) is known to affect an API variable (U), potentially violating the exclusion restriction. Any other independent variable (NC) that affects this API variable as well can serve as an NCI. In this example, the IV is independent of the NCI ($Z \perp\!\!\!\perp NC$). Therefore an unconditional test can be utilized. By contrast, when 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. Previous literature on negative control tests has mostly focused on causal analysis without an IV, which we do not discuss in this paper. In non-IV settings, there is no exclusion restriction and therefore independence tests between a negative control and the outcome are always done conditionally.[7]

[7] The analog of NCI in non-IV settings is negative control exposure (NCE). NCE tests are always conducted conditional on the exposure.

2.4 Control Variables and Functional Forms

In many cases, the IV is believed to be exogenous only conditional on certain control variables (e.g., assignment of judges is quasi-random only within time and location; Kling, 2006). Let C be the vector of controls. Similar to the case without controls, outcome independence and exclusion restriction together imply $Z \perp\!\!\!\perp Y(x)|\ C$. In Appendix D, we present the theory of alternative path variables and negative controls when controls are included. When the IV is presumably valid only conditional on a vector of control variables C, an NCO test is a test for the null hypothesis

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

Similarly, for NCIs the null hypothesis is

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

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 assumptions.

NCO tests typically adopt the functional form assumption on how the IV depends on the control variables. With some abuse of notation to avoid clutter, let C also denote the set of controls in a 2SLS specification.[8] Blandhol et al. (2022) show that 2SLS requires the following linearity assumption to satisfy their definition of a *causal estimand*.[9]

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 satisfies the NCO assumption. Let*

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

[8] 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.

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

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.

The proof is given in Appendix D.2.

Turning to the NCI tests, such tests typically adopt the functional form assumptions on 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, $\mathbb{E}[Y|Z, C] = \theta_Z Z + \theta_C' C$.*

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 exogeneity and CSRF.

Corollary 2. *Assume that the random variable NC satisfies the NCI assumption. 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.

The proof is given in Appendix D.2. Corollaries 1 and 2 imply that in the tests discussed, the null hypothesis can be rejected in exogenous IV designs. 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 by using methods other than 2SLS (Blandhol et al., 2022). For NCI tests, the null can be rejected because CSRF is violated. However, CSRF is not a necessary assumption for 2SLS analysis, implying that testing this assumption could reject perfectly valid IV designs.

For example, an NCI test can reject a design where the IV is randomly assigned due to CSRF violation. Random assignment guarantees that outcome independence and rich covariates hold. Assume that the exclusion restriction holds as well and therefore the design is valid. 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 a case, an NCI test could reject the null hypothesis in (5), even though the design is valid.

3 Negative Control Test Procedures

This section discusses conditional independence tests, which our theory shows are essential for many negative control tests in IV designs.[10] We first review the methods currently employed in economics research for conducting conditional independence tests with negative controls. We then introduce and evaluate several underutilized testing methods that can enrich the toolkit available to researchers. Conditional independence, especially with continuous controls, has no established one-size-fits-all solution (Shah and Peters, 2020). We use simulations to demonstrate the trade-offs between tests that rely on strong functional form assumptions and more flexible tests that require large sample sizes. Practical recommendations for selecting appropriate statistical tests are provided in Section 4.

3.1 Pseudo-Outcome and Pseudo-IV Tests

As discussed in Section 1, the most commonly employed falsification tests in practice are *pseudo-outcome* tests. These tests estimate the original reduced form equation while replacing the outcome with a pseudo-outcome (e.g., past outcomes). That is, a pseudo-outcome analysis estimates the model

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

and examines the estimated coefficient of the IV in this model (e.g., by testing $H_0 : \beta_Z = 0$).[11]

The following result establishes that when the pseudo-outcome is a valid NCO, this practice can be used to jointly examine IV exogeneity and rich covariates.

Corollary 3. *Assume that NC satisfies the NCO assumption. Let*

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

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

The proof is given in Appendix D.2.

Unlike pseudo-outcome analysis, *pseudo-IV* analysis, which is the common approach for NCI tests, is not always a valid test for IV exogeneity. In its simplest form, pseudo-IV analysis

[10] Whenever control variables are included in the design, negative control tests examine conditional independence (Equations (2) and (3)). Moreover, NCI tests typically require conditional independence tests even without control variables. NCO tests without controls and some NCI tests without controls require unconditional independence testing (Theorems 1 and 3). Such independence is testable using various established statistical tests (e.g., Székely et al., 2007; Heller et al., 2013).

[11] The chosen testing procedure for H_0 may vary according to the assumptions on ϵ_{NC}.

substitutes the original IV with a similar but ostensibly unrelated variable (the pseudo-IV) in the reduced form regression and assesses the independence between the pseudo-IV and the outcome. Such a test is not informative about IV exogeneity as the pseudo-IV is not independent of the original IV. Therefore, the original IV must be controlled for (Theorem 2). Notably, this is not consistently done in empirical studies, as discussed in Section 1.

A pseudo-IV analysis that adjusts for the IV and the controls may use the linear regression

$$Y = \theta_Z Z + \theta_C' C + \theta_{NC} NC + \epsilon_Y, \tag{7}$$

where NC is an NCI. The pseudo-IV test then examines the null hypothesis $H_0 : \theta_{NC} = 0$.[12]

Despite the adjustment, this pseudo-IV analysis could still reject the null hypothesis for valid IV designs. Corollary 2 implies that even when the IV is exogenous, the null could still be rejected due to a violation of CSRF. As discussed in Section 2.4, the CSRF assumption is not necessary for causal identification. Therefore, it is possible that an adjusted pseudo-IV test would reject the null hypothesis, even though the IV design can identify a causal estimand with 2SLS. In Sections 3.3 and 3.4, we discuss tests for IV exogeneity that do not rely on or test additional functional form assumptions.

When multiple negative controls are available, multiple pseudo-outcome or pseudo-IV analyses can be conducted separately for each negative control. This procedure requires multiple hypothesis-testing corrections, such as Bonferroni correction. Alternatively, we now turn to discuss procedures that perform a joint test for multiple negative controls. A comparison between the different approaches is part of our simulation study (Section 3.5).

3.2 Parametric Conditional Independence Tests

When multiple negative controls are available, one may use a joint test to combine information from multiple negative controls.[13] Let NC be a vector of NCOs, and consider the following linear model for Z:

$$Z = \gamma' C + \gamma_{NC}' NC + \epsilon_Z. \tag{8}$$

An F-test can test the null hypothesis $H_0 : \gamma_{NC}' = 0'$ (under the standard assumptions). Alternatively, a Wald test might be preferred when robust or clustered standard errors are used. This procedure jointly tests IV exogeneity and rich covariates (Corollary 1).

Similarly, with multiple NCIs, an F-test (or other tests for a vector of parameters) can be used in an adjusted pseudo-IV analysis. Using a vector of NCIs in Equation (7), researchers

[12] The testing procedure is determined according to additional assumptions on ϵ_Y, as before.

[13] In most applications, an association with a vector of negative controls implies an association with at least one of its components. Appendix C.10 provides a knife-edge counterexample of a case where a vector of negative controls is not a negative control. However, a small perturbation to the parameter values in this example would reverse the conclusion.

can carry out a valid statistical test for the hypothesis that the coefficient vector of the NCIs is zero ($H_0 : \theta'_{NC} = 0'$). This procedure jointly tests IV exogeneity and the CSRF assumption (Corollary 2). Therefore, it could reject valid IV designs as well.

3.3 Semi-Parametric Conditional Independence Tests

In order to test IV exogeneity without testing functional form assumptions, researchers may opt for semi- or non-parametric tests. As an alternative to the parametric tests we have discussed so far, consider, for example, the use of general additive models (GAMs) (Wood, 2006). For NCO tests, one such specification assumes that

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

where C, NC are vectors of controls and NCOs indexed by j and k, respectively; f_j, g_k are smooth functions allowed to be different for different variables; and ϵ_Z is an independent normally distributed error term. Under this model, Z is composed of additive smooth functions that are typically estimated using splines. A Wald test for the spline coefficients can be used to test the null hypothesis that $g_k = 0$ for all k.[14] A GAM test can be similarly implemented for NCI tests, by adding smooth functions to Equation (7). Interactions can be accommodated by including additional smooth functions of product terms.

A GAM model can also be used to test the rich covariates assumptions, while still allowing for a nonlinear association between the NCOs and the IV. Such a model is useful when one uses 2SLS (which requires rich covariates) while suspecting a strong nonlinear association between the NCO and the IV. In such cases, the following model can be used:

$$Z = \phi(C, NC) + \epsilon_Z = \gamma' C + \sum_k g_k(NC_k) + \epsilon_Z. \tag{10}$$

As with model (9), the NCO test is for the hypothesis that $g_k = 0$ for all k. When the hypothesis is rejected, g_k may capture a residual nonlinear association between C and Z, not captured by γ, which violates rich covariates (due to an association between C and NC).

3.4 Non-Parametric Conditional Independence Tests

With a large number of negative controls or a large dataset that allows one to estimate flexible non-parametric models, researchers can implement a conditional independence test using invariant target prediction (Heinze-Deml et al., 2018). In the context of an NCO test,

[14] See Section 6.12.1 in Wood (2006) for technical details and conditions for test validity.

this method involves using a prediction algorithm to predict the IV twice: once using only the control variables and once using both the NCOs and the control variables. Under the null hypotheis ($Z \perp\!\!\!\perp NC|C$), the out-of-sample performance of both predictions should be similar, as the NCOs should not have any additional predictive power if they are conditionally independent of the IV. A similar approach can be constructed for NCI tests (while using the IV in both prediction algorithms). For a large enough sample size, various prediction algorithms, including machine-learning methods, can be employed in this procedure.[15]

Similar to the other methods we have discussed so far, invariant target prediction typically focuses on mean independence. However, more general approaches go beyond mean independence by using kernel measures of conditional independence (Fukumizu et al., 2007; Zhang et al., 2011; Strobl et al., 2019).

3.5 Simulations

This section summarizes two simulation studies comparing the negative control tests discussed earlier. For simplicity, we focus on NCO tests. See Appendix E for details about the data-generating processes (DGP) and specific parameter values for each study.

In each simulated sample, we test the null hypothesis $H_0 : \mathbb{E}[Z|C, NC] = \mathbb{E}[Z|C]$ using the following tests: multiple pseudo-outcome regressions (Equation (6)) with Bonferroni correction for multiple hypotheses, a single (multivariate) linear regression with an F-test (Equation (8)), and Wald tests from GAMs with and without smooth terms for the control variables (Equations (9) and (10)).[16] We set the desired type-I error to 5%.

In the first study, we simulate a violation of outcome independence. Panel A of Figure 3 summarizes the results. We separately examine a scenario with a linear association between the IV and the NCOs (left panel) and a highly nonlinear scenario. For each scenario, we report the null rejection rate across simulated datasets as a function of the IV–NCO relationship. In the linear scenario, the F-test outperformed other methods, especially when the relationship between the NCOs and the IV was weak. The multiple pseudo-outcome regressions surpassed GAMs under a linear relationship but had lower power than the F-test. Conversely, under strong nonlinearity, only the GAM tests had satisfactory power.

In the second study, we consider test performances when IV exogeneity holds but rich covariates is violated. That is, the conditional expectation of the IV is not linear in the control variables. We report the rejection rate as a function of the strength of this nonlinear relationship (Panel B). Tests assuming a linear association between the IV and the control variables rejected the null when the nonlinearity in the IV–controls association was

[15] For example, Abramitzky et al. (2023) use LASSO to predict the IV with the controls and NCOs.

[16] We also considered a Wald test from the single linear regression; the results were nearly identical to those obtained by the F-test and hence are not reported.

substantial (i.e., a substantial violation of rich covariates). The flexible GAM rarely rejected the null, indicating that it is not testing the rich covariates assumption. Therefore, as expected from the theory, linear tests assess both functional form and IV exogeneity, whereas nonlinear tests primarily evaluate IV exogeneity.

These simulations suggest that linear tests are preferred when the associations are closer to linear, the sample size is small, or when examining functional form assumptions is desired. Conversely, nonlinear tests are preferred when nonlinear associations are suspected, the sample size is large, or when the researcher wants to focus solely on IV exogeneity.

4 Practical Guidance

This section offers guidelines for implementing negative control tests. We recommend that researchers follow three steps, summarized in Appendix Figure A1. First, use domain knowledge to identify suitable negative controls. Examples are discussed in Section 4.1. Second, choose an appropriate statistical test based on sample size and functional form assumptions, as detailed in Section 4.2. Third, interpret the result and conduct further diagnostics if the test rejects the null, as discussed in Section 4.3.

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* (AER) with publicly posted replication data. We use Autor et al. (2013) and Deming (2014) to discuss NCO tests and Ashraf and Galor (2013) and Nunn and Qian (2014) to discuss NCI tests.[17] Appendix Table A1 summarizes these papers' IV designs and the negative controls they used in falsification tests. Appendix F provides additional details on our analyses.

4.1 Choosing Negative Controls

In our survey of recent literature in Section 1, we found that researchers often use only one, or very few negative controls. This may raise the concern that valid and informative negative controls are sometimes left unused. Guided by our theory, we found additional negative controls in the replication data of some of the papers we analyzed (Column 6 of Appendix Table A1).

We now discuss different types of negative controls, both commonly used in practice and novel ones suggested by our theoretical framework.

[17] 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 et al. (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 4.3.

4.1.1 Common Types of Negative Control Outcomes

Predetermined Variables. Predetermined variables are frequently used in NCO tests. Common examples include lagged outcomes and demographic factors such as gender, race, and age. Autor et al. (2013) used predetermined local labor market manufacturing employment to evaluate a shift-share IV for commute-zone exposure to imports. Following the same logic, we found many additional predetermined variables in the original paper's replication data that could serve as NCOs (e.g., past unemployment in the local labor market). Similarly, we found multiple predetermined variables in the replication data from Deming (2014). These are all NCOs as the IV is based on a random school lottery.

Predetermined variables can proxy for APO variables that violate outcome independence if they affect the IV. In many cases, researchers choose to use predetermined variables as NCOs while only vaguely knowing which APO variable they proxy for. The NCO assumption requires that if the NCO is associated with the IV, the association with the IV is through some APO variable.

However, not every predetermined variable is a valid NCO. Certain predetermined variables may influence the IV, even if the IV is exogenous, thus violating the NCO assumption. For example, when the IV is the child's quarter of birth (Angrist and Krueger, 1991), the parents' quarter of marriage is not a valid NCO as it is likely affecting it. Such concerns are less relevant when the IV is based on an alleged randomization (e.g., school lotteries).

IV Leads and Lags. Certain IVs are predicated on serendipitous or chance occurrences ("strokes of luck"). In these cases, the IV should not be correlated with future or past measurements of the IV, which can serve as NCOs. For example, Jäger and Heining (2022) use a worker's premature death as an IV for employee turnover, under the assumption that such deaths are random across firms. One potential concern in this example is that premature deaths may result from riskier conditions in the firm, which could directly impact wages (the outcome). To rule this out, Jäger and Heining use subsequent premature deaths in the same firm as an NCO that proxies for the APO variable (riskier conditions). A recurring pattern of premature deaths could cast doubt on the assumption that such deaths are random across firms. They find no correlation of premature deaths within a firm over time. The NCO assumption here stipulates that subsequent IV measurements should not be autocorrelated.

Alternative Outcomes. Unrelated or alternative outcomes can also serve as NCOs for two different types of APO variables. First, when the APO variable is potentially violating outcome independence (potentially affecting the IV), alternative outcomes can serve as NCOs, provided they are unaffected by the IV. Chetty et al. (2014) evaluate middle-school teacher value-added measures using a movers design, leveraging teachers' movements between schools. Here, the APO variable of concern is the unobserved changes in school quality. For example, high-quality teachers may tend to move to schools experiencing simultaneous improvements in student quality. To evaluate this threat, Chetty et al. 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 movers design's validity.

The second type of APO variable 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 Figure 1B). For example, Angrist and Evans (1998) use the sex composition of the first two children as an IV for the total number of children. One potential concern is that same-sex sibship could reduce housing expenditures because of hand-me-downs, which could then affect female labor supply decisions (the outcome). To test this, Rosenzweig and Wolpin (2000) show that same-sex sibship also affects an alternative outcome, clothing expenditures. Clothing expenditures is an NCO, which proxies for overall expenditures (the APO variable). In this case, the NCO assumption does not hold if the IV affects the candidate NCO through the outcome, i.e. if female labor supply affects household expenditure on clothing.

4.1.2 Common Types of Negative Control Instruments

Pseudo-IVs. Researchers often choose NCIs that are similar to the IV but are presumed not to influence the treatment variable. These pseudo-IVs usually share many similarities to the IV and are thus likely to be correlated with the API variable. For example, as discussed in the introduction (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 as NCIs. These variables are similar, as they are both affected by the same API variables such as weather conditions.

IV Leads. Future instances of the IV (IV leads) can often serve as effective NCIs. 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 (2021) then ascertains that predicted future cluster size is not additionally correlated with productivity. This relies on the fact that IV leads, which are based on events that occur after the outcome, cannot influence it. As with pseudo-IVs, 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 IV to be autocorrelated. When the IVs are expected to be autocorrelated (as in Moretti, 2021) an NCI strategy can be used. When the IVs are presumably uncorrelated, and NCO strategy can be applied, as discussed in the previous section.

4.1.3 Novel Types of Negative Control Instruments

Our theory can highlight potential new types of NCIs beyond those commonly used today.

Causes of the IV. Variables that causally affect the IV can serve as effective NCIs for assessing possible violations of outcome independence. Moreover, these NCIs do not require the researcher to take a stand on what is the potential violation. Panel B of Figure 2 illustrates how such NCIs works. For example, if the IV is wheat production (Nunn and Qian, 2014), then observed weather conditions that affect wheat production qualify as a valid NCI. In such scenarios, both the API variable and the NCI influence the IV, leading to a correlation between the API variable and the NCI, conditional on the IV. The IV is what is known in the literature of DAGs as a *collider* (Pearl, 2009). If the NCI is correlated with the outcome conditional on the IV, this implies that there is a path between the NCI and the outcome, through the IV and the API variable. Therefore, the API affects the outcome, and the IV design is not exogenous.

IV Side Effect Proxies. An IV that affects an API variable in addition to its effect on the treatment may violate the exclusion restriction if this API variable also affects the outcome. Therefore, proxies for such API variables can serve as NCIs. Panel D of Figure 1 illustrates a scenario where the IV influences the NCI through the API variable, and therefore the NCI is an alternative outcome. Panel A of Figure 2 presents another case where the NCI and the IV both affect the API variable. These “side effect” proxy variables must not be associated with the outcome other than through the IV and the API variable, as any other association would violate the NCI assumption.

For example, some papers use a regression discontinuity to identify an effect of a policy that only applies above a certain cutoff (e.g., municipalities above a certain population size). However, the same cutoff is sometimes used for more than one policy. If the additional policy does not affect the outcome, the design is still valid. To test this, any proxy for participation in the additional policy can be used as an NCI.

4.1.4 Power Considerations

Power considerations suggest excluding variables that meet a negative control assumption but have weak associations with the alternative path variables, as they can lower test power.

This mirrors the effect of irrelevant control variables in OLS. This issue is especially acute in pseudo-IV NCI tests, as pseudo-IVs can be strongly correlated with the IV but only weakly correlated with the API variable conditional on the IV.

4.2 Choosing a Statistical Test

When implementing negative control tests, researchers can choose from a variety of statistical tests for (conditional) independence between the negative control and the IV or the outcome. Different testing methods were discussed in Section 3. The appropriate statistical test depends on three primary considerations: the design assumptions under scrutiny, the expected functional relationship between the negative control and either the IV or the outcome, and statistical power.

NCO Tests. When a 2SLS specification with controls is used for estimation, an NCO test that is linear in the controls is a leading option. Specifically, we recommend estimating (8) and using an F-test or a Wald test for the hypothesis that the NCO coefficients equal zero. Such tests evaluate not only IV exogeneity but also the necessary rich covariates assumption (Assumption 3). Moreover, this test can aggregate information from multiple NCOs jointly. For a large enough sample size, a GAM model as in (10) can allow for nonlinear associations with the NCO while still testing the rich covariates assumption.

Alternatively, researchers might opt for a pseudo-outcome test based on (6). This test uses the same inferential framework as the original study. Therefore, it can expose errors in the inference method as well (Eggers et al., 2023). However, it demands accounting for multiple hypothesis testing when multiple NCOs are available.

When the main analysis does not use 2SLS (e.g., in quantile regressions as in Chernozhukov and Hansen, 2008), and the sample size is sufficiently large, nonlinear tests are preferred. This includes semi-parametric methods like the GAM in (9), or other methods discussed in Sections 3.3 and 3.4. These tests can capture more complex associations that would be missed in linear tests. However, their flexibility implies larger sample size requirements, as demonstrated in our simulations (Section 3.5).

NCI Tests. For NCI tests, researchers should first consider whether they need to control for the IV in their test according to whether the NCI is independent of the IV (Section 2.3.3). In most cases, controlling for the IV is required. One notable exception is when the API variable represents a violation of the exclusion restriction, and the NCI affects the API variable (Panel A of Figure 2).

Researchers should then determine which statistical test to use. The commonly used approach that adds the NCIs to the reduced form equation is often problematic, as its validity depends on the CSRF assumption (Assumption 4). At face value, this strong parametric assumption is attractive as it improves the statistical power when the sample size is small. However, this assumption is not necessary and is often violated, which could lead to the rejection of valid IV designs. With a sufficiently large sample size, researchers should instead use a semi- or non-parametric test, as discussed above.

Examples of Applications. Table 3 presents the results of all the above-discussed NCO tests using the data from Autor et al. (2013) and Deming (2014). Column (2) shows that in both papers, the null hypothesis is not rejected when only one NCO is used.[18] By contrast, Column (3) shows that the null hypothesis is rejected when multiple NCOs are used (with Bonferroni correction for multiple hypothesis-testing). Column (4) shows that the null is also rejected when a joint F-test is used. These results underscore that adding additional NCOs—using existing data and guided by theory—can yield more powerful tests.[19] In Columns (5) and (6) we implement a GAM test with and without assuming linearity of the control variables. As expected, the GAM test is less effective in smaller samples.

Table 4 presents results from the above NCI tests using data from Nunn and Qian (2014) and Ashraf and Galor (2013). Columns (1) and (2) show the results of a pseudo-IV test with and without conditioning on the IV, using one of the NCIs from the above-cited papers. In both contexts, conditioning on the IV is required as the NCI is likely associated with the outcome through the IV.[20] For Nunn and Qian, we find that the null hypothesis is only rejected when conditioning on the IV is omitted, indicating that neglecting to control for the IV would have led to a false rejection of their IV design. Using data from Ashraf and Galor, the null is not rejected, even without controlling for the IV, potentially due to insufficient statistical power. In Columns (3) and (4) we show the null is not rejected when using the full set of NCIs from the above-cited papers in either multiple pseudo-IV tests with Bonferroni correction or in a joint F-test. The small sample size compared to the large number of control variables does not allow proper estimation of GAM models in either paper.[21]

4.3 Interpreting the Test Results

Rejection of the Null. Rejecting the null hypothesis in a parametric linear negative control test, such as an F-test, could suggest violations of either IV exogeneity or linearity

[18] In Column (1) we show the results from a replication of the original NCO test conducted by Autor et al. (2013). They find a significant association between the IV and the lagged outcome, with a sign that is the opposite of the sign in the main analysis. This association becomes insignificant when all control variables are included.

[19] In Section 4.3 we use additional analysis to detect a fixable problem in the IV design in Deming (2014).

[20] While not explicitly stated in Nunn and Qian (2014), their code reveals conditioning on the IV. Ashraf and Galor (2013) did not condition on the IV.

[21] We estimate a GAM model with linear controls for Nunn and Qian (2014). See Appendix F.3 for details.

(rich covariates for NCO, CSRF for NCI). In such cases, researchers can further explore the reason for the rejection by testing the linearity assumption directly.[22] For a large enough sample size, researchers can also use semi- or non-parametric tests, which rely on fewer or no functional form assumptions and focus on IV exogeneity.

When multiple negative controls are used, researchers may want to explore which negative controls are driving the rejection of the null. To do so, researchers can plot the strengths of the correlations for each negative control with the IV and the outcome on a two-dimensional scatter plot. This exercise can help identify negative controls with high predictive power with regard to both the IV and outcome, which may suggest the existence of a related alternative path. Appendix Figure A2 displays such diagnostics for Deming (2014) who uses school lotteries to evaluate school value-added measures (see Appendix Table A1 for details). We plot each NCO's correlation with the IV (vertical axis) against its correlation with the outcome (horizontal axis). The NCO with the strongest correlation with the IV is the value added of the neighborhood school. This is because the IV is constructed from interactions of the lottery results with neighborhood school VA. In this case, the post-analysis exercise identifies a fixable problem in the construction of the IV. Using the original lottery results, we get noisier estimates, however, we cannot rule out that the main results are unchanged.

One caveat for this exercise is that the correlation with negative control may not always reflect how strong the alternative path is. Since negative controls are just proxies for alternative path variables, their correlation with the IV or the outcome depends on the strength of their correlation to the alternative path variable. Therefore, a negative control might have a weak correlation with the IV or outcome, even if the association between the IV and outcome through the alternative path is strong, or vice versa.

Non-Rejection of the Null. As discussed in Section 2, if the null hypothesis is not rejected, the IV might still not be exogenous. First, IV exogeneity may be invalidated by alternative path variables not captured by the NCO or NCI used for the test. For example, a quasi-random allocation to teachers that is found to be uncorrelated with students' neighborhoods could still be associated with students' abilities within neighborhoods. Second, invalid IVs could pass the test due to lack of power.

5 Bias Correction

So far we have only discussed negative controls in the context of bias detection. In this section, we turn to discuss the possibility of using negative controls for bias correction.

[22] E.g., by using a RESET test with control variables only, without negative controls (Ramsey, 1969).

One common, albeit imperfect, approach to bias correction is to use the negative controls as control variables. For example, Autor et al. (2013) show that their results are robust for controlling for lagged commuting zones characteristics, which could also serve as NCOs. However, as discussed in Section 2, negative controls are often only imperfect proxies for the threat to the identification. Controlling for an imperfect proxy may lower the bias but does not eliminate it.

In non-IV settings, negative controls are frequently used for bias correction. For example, the difference-in-differences (DiD) estimator can be seen as a form of bias correction using an NCO. In a two-period DiD, the NCO is the preperiod outcome (Sofer et al., 2016). The average difference in the NCO between treated and untreated units is an estimate of the confounding bias. Hence, by subtracting this difference, the bias is corrected (under the standard DiD assumptions; Roth et al., 2023). Moreover, when the NCO does not directly quantify the bias, another type of negative control can be used to scale the coefficient on the NCO such that it can be subtracted (Shi et al., 2020).[23] In this section, we consider a simple case where the bias can be directly subtracted, making scaling unnecessary.

For this section, we assume the following simple IV setting. Let Z_1, X_1, and Y_1 be the IV, treatment, and outcome, respectively, measured in a period indexed by 1. Assume that an NCO, Y_0, is also observed. We refer to Y_0 as the lagged outcome (indexed here by 0); however, it could also represent any alternative pseudo-outcome.[24] In this section, we focus on the simple case where both the IV (Z_1) and the treatment (X_1) are binary.

Similar to the canonical DiD model (without an IV), we assume that neither the treatment nor the IV has been initiated in the preperiod ($Z_0 = X_0 = 0$). This is a typical scenario in analyses of new policies. In Appendix D.3, we discuss alternative preperiod scenarios, which are less likely to occur in practice but require more plausible assumptions for bias correction.

Suppose that a pseudo-outcome test reveals a correlation between the IV and the NCO. We refer to the pseudo-effect as the difference in the average NCO between IV-treated and IV-untreated units, $\mathbb{E}[Y_0|Z_1 = 1] - \mathbb{E}[Y_0|Z_1 = 0]$. We assume that outcome independence (Assumption 1) does not hold and, therefore, the pseudo-effect is different from zero. We investigate whether the pseudo-effect on the NCO can be informative on the bias.

Specifically, we study under what conditions the pseudo-effect can be simply subtracted to correct the bias. To this end, consider the following *difference-in-Wald* (DiW) estimator.

Definition 6 (Difference in Wald)**.**

$$\tau^{DiW} = \frac{\mathbb{E}\left[Y_1 - Y_0|Z_1 = 1\right] - \mathbb{E}\left[Y_1 - Y_0|Z_1 = 0\right]}{\mathbb{E}\left[X_1|Z_1 = 1\right] - \mathbb{E}\left[X_1|Z_1 = 0\right]}.$$

[23] The double-negative control design is part of a more general approach to use proxies to study causal effects. This approach has recently been termed “causal proximal learning” (Tchetgen Tchetgen et al., 2020).

[24] For example, if Y_1 is test scores in some subject, Y_0 could denote test scores in an alternative subject.

Additional assumptions are needed for τ^{DiW} to identify a causal effect. First, assume that treatment independence ($Z_1 \perp\!\!\!\perp X_1(z)$) and relevance ($\mathbb{E}[X_1(1) - X_1(0)] \neq 0$) both hold.[25] Assume also that the exclusion restriction (Assumption 2) holds.

Second, we make assumptions analogous to DiD assumptions. Since the IV and the treatment have not been initiated in period 0, we can write $Y_t(z_1, x_1)$ for the potential outcomes at time $t = 0, 1$ when intervening on the IV or the treatment in period 1.[26]

Assumption 5 (DiD assumptions for IV)**.**

1. No Anticipation. *For all $z_1, x_1, \tilde{z}_1, \tilde{x}_1$, $Y_0(z_1, x_1) = Y_0(\tilde{z}_1, \tilde{x}_1)$.*

2. Parallel Trends. $\mathbb{E}[Y_1(0,0) - Y_0(0,0)|Z_1 = 1] = \mathbb{E}[Y_1(0,0) - Y_0(0,0)|Z_1 = 0]$.

The first condition implies that the lagged outcome Y_0 is not affected by the future treatment or IV (as in the canonical DiD model; Roth et al., 2023). A violation of this condition might imply that Y_0 does not satisfy the NCO assumption. The second condition generalizes the DiD parallel-trends assumption for the IV setting. It is the counterfactual statement that in the absence of the IV and the treatment, the trend in the mean outcome would have been the same for the IV-treated and IV-untreated observations.

The DiW estimator requires an additional assumption to identify a causal effect. One option is to assume a homogeneous treatment effect. Since we assumed that the exclusion restriction holds, we can write the treatment effect as $Y_1(z_1, 1) - Y_1(z_1, 0) = Y_1(1) - Y_1(0)$.

Assumption 6 (Homogeneous treatment effect)**.** *Assume that for every observation, $Y_1(1) - Y_1(0) = \tau$, for some constant τ.*

The following theorem establishes that τ is identified by the DiW estimator under the above assumptions. Hence, the pseudo-effect on the NCO can be used for bias correction.

Theorem 4. *Assume that treatment independence, exclusion restriction, Assumptions 5 and 6 hold. The difference-in-Wald estimator identifies the causal effect. Namely, $\tau^{DiW} = \tau$.*

When the heterogeneity in the treatment effect is substantial, the bias in τ^{DiW} can be arbitrarily large. In Appendix D.3 we show that in cases where the treatment can be received together only with the IV ($X_1(0) = 0$), the homogeneous treatment effect assumption is not required. An example is when the IV is a voucher to participate in the treatment, and participation is impossible without the voucher. In these cases, τ^{DiW} equals the average treatment effect for the treated compliers (Angrist et al., 1996).

[25] In some contexts, bias in the reduced form implies that treatment independence is also violated. Without treatment independence, the first-stage estimates would require correction as well.

[26] Similar to Roth et al. (2023), we write $Y_t(z_1, x_1) = Y_t(0, 0, z_1, x_1)$, where the first two arguments in $Y_t(z_0, x_0, z_1, x_1)$ denote the IV and the treatment in period 0.

In Appendix D.3, we discuss two alternative and less common scenarios in which the DiW estimator is unbiased under more plausible assumptions. We show that under different preperiod regimes (i.e., different assignments of Z_0, X_0), treatment effect homogeneity is not required, and violation of the exclusion restriction can also be corrected. However, such scenarios are less frequent in practical applications.

6 Conclusion

This paper develops a theoretical framework for negative control tests for IV designs. Our results uncover the underlying assumptions behind practices that are frequently applied and give formal justifications for common intuitions. Moreover, we highlight four key findings that, in our view, are not commonly known, and could have practical implications on how negative control tests are used in practice.

First, most NCI tests are implemented incorrectly as they do not control for the original IV. This could lead to the rejection of valid IV designs. Second, common negative control tests are not only testing the exogeneity of the IV but also testing functional form assumptions, which are replaceable and sometimes unnecessary. Third, our theory 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, we believe that in many cases, negative control variables are readily available in researchers' data and should be used to construct more powerful negative control tests. Finally, we find that under stronger assumptions, negative controls can be used not only to detect but also to correct biases in IV designs. Ultimately, we believe this work will contribute to a more systematic and effective use of negative control falsification tests for evaluating and amending IV designs.

References

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

Abramitzky, Ran, Philipp Ager, Leah Boustan, Elior Cohen, and Casper W. Hansen, "The effect of immigration restrictions on local labor markets: Lessons from the 1920s border closure," *American Economic Journal: Applied Economics*, 2023, *15* (1), 164–191.

Acemoglu, Daron, Giuseppe De Feo, and Giacomo Davide De Luca, "Weak states: Causes and consequences of the Sicilian Mafia," *Review of Economic Studies*, 2020, *87* (2), 537–581.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

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

Angrist, Joshua D. and William N. Evans, "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 1998, *88* (3), 450–477.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

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

Athey, Susan and Guido W. Imbens, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic Perspectives*, 2017, *31* (2), 3–32.

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

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

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

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Chyn, Eric, Brigham Frandsen, and Emily C Leslie, "Examiner and Judge Designs in Economics: A Practitioner's Guide," *NBER Working Paper 32348*, 2024.

Davies, Neil M., Kyla H. Thomas, Amy E. Taylor, Gemma M.J. Taylor, Richard M. Martin, Marcus R. Munafò, and Frank Windmeijer, "How to compare instrumental variable and conventional regression analyses using negative controls and bias plots," *International Journal of Epidemiology*, 2017, *46* (6), 2067–2077.

De Giorgi, Giacomo, Anders Frederiksen, and Luigi Pistaferri, "Consumption network effects," *Review of Economic Studies*, 2020, *87* (1), 130–163.

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

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

Doyle, Joseph J., John A. Graves, Jonathan Gruber, and Samuel A. Kleiner, "Measuring returns to hospital care: Evidence from ambulance referral patterns," *Journal of Political Economy*, 2015, *123* (1), 170–214.

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

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

Fukumizu, Kenji, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf, "Kernel measures of conditional dependence," *Advances in Neural Information Processing Systems*, 2007, *20.*

Glymour, M Maria, Eric J Tchetgen Tchetgen, and James M Robins, "Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions," *American journal of epidemiology*, 2012, *175* (4), 332–339.

Guidetti, Bruna, Paula Pereda, and Edson Severnini, "'Placebo tests' for the impacts of air pollution on health: The challenge of limited health care infrastructure," *AEA Papers and Proceedings*, 2021, *111*, 371–75.

Heinze-Deml, Christina, Jonas Peters, and Nicolai Meinshausen, "Invariant causal prediction for nonlinear models," *Journal of Causal Inference*, 2018, *6* (2), 1–35.

Heller, Ruth, Yair Heller, and Malka Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, 2013, *100* (2), 503–510.

Huber, Martin and Giovanni Mellace, "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 2015, *97* (2), 398–411.

Imbens, Guido W., "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics," *Journal of Economic Literature*, 2020, *58* (4), 1129–79.

Jäger, Simon and Jörg Heining, "How substitutable are workers? Evidence from worker deaths," *NBER Working Paper 30629*, 2022.

Keele, Luke, Qingyuan Zhao, Rachel R Kelz, and Dylan Small, "Falsification tests for instrumental variable designs with an application to tendency to operate," *Medical care*, 2019, *57* (2), 167–171.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad, "Field of study, earnings, and self-selection," *Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

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

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

Lipsitch, Marc, Eric Tchetgen Tchetgen, and Ted Cohen, "Negative controls: A tool for detecting confounding and bias in observational studies," *Epidemiology (Cambridge, Mass.)*, 2010, *21* (3), 383.

Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott, "Do political protests matter? Evidence from the Tea Party movement," *Quarterly Journal of Economics*, 2013, *128* (4), 1633–1685.

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

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

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

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

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

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

Ramsey, James Bernard, "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1969, *31* (2), 350–371.

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

Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe, "What's trending in difference-in-differences? A synthesis of the recent econometrics literature," *Journal of Econometrics*, 2023, *235* (2), 2218–2244.

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

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

Sofer, Tamar, David B. Richardson, Elena Colicino, Joel Schwartz, and Eric J. Tchetgen Tchetgen, "On negative outcome control of unobserved confounding as a generalization of difference-in-differences," *Statistical Science*, 2016, *31* (3), 348–361.

Strobl, Eric V., Kun Zhang, and Shyam Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," *Journal of Causal Inference*, 2019, *7* (1), 20180017.

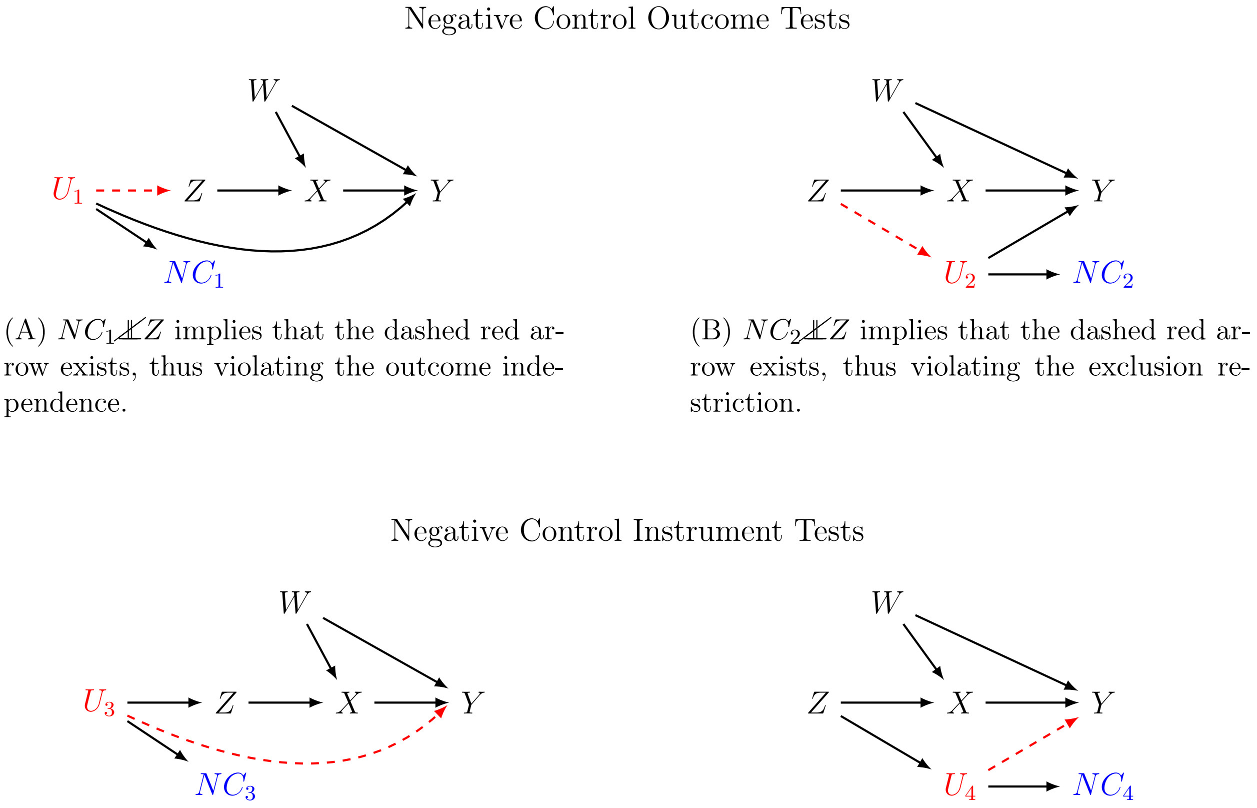
Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov, "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, 2007, *35* (6), 2769–2794.

Tchetgen Tchetgen, Eric J., Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao, "An introduction to proximal causal learning," *arXiv preprint arXiv:2009.10982*, 2020.

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

Zhang, Kun, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 804–813.

Figure 1: Negative Control Falsification Tests: Graphical Illustrations



Negative Control Outcome Tests

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

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

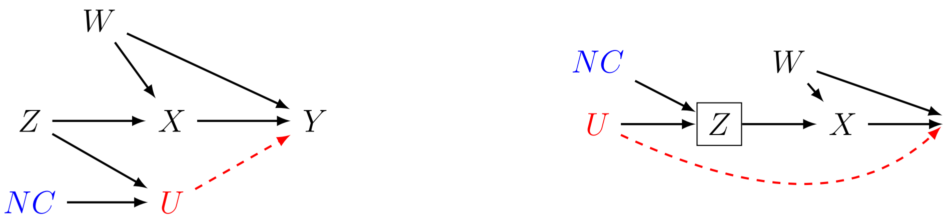
Negative Control Instrument Tests

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

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

Notes: The figure illustrates how negative control tests assess the exogeneity 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 red arrows exist. The top panels (A and B) depict negative control outcome tests. IV exogeneity 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 Z is independent of NC_i. 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 $NC \perp Y|Z$.

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

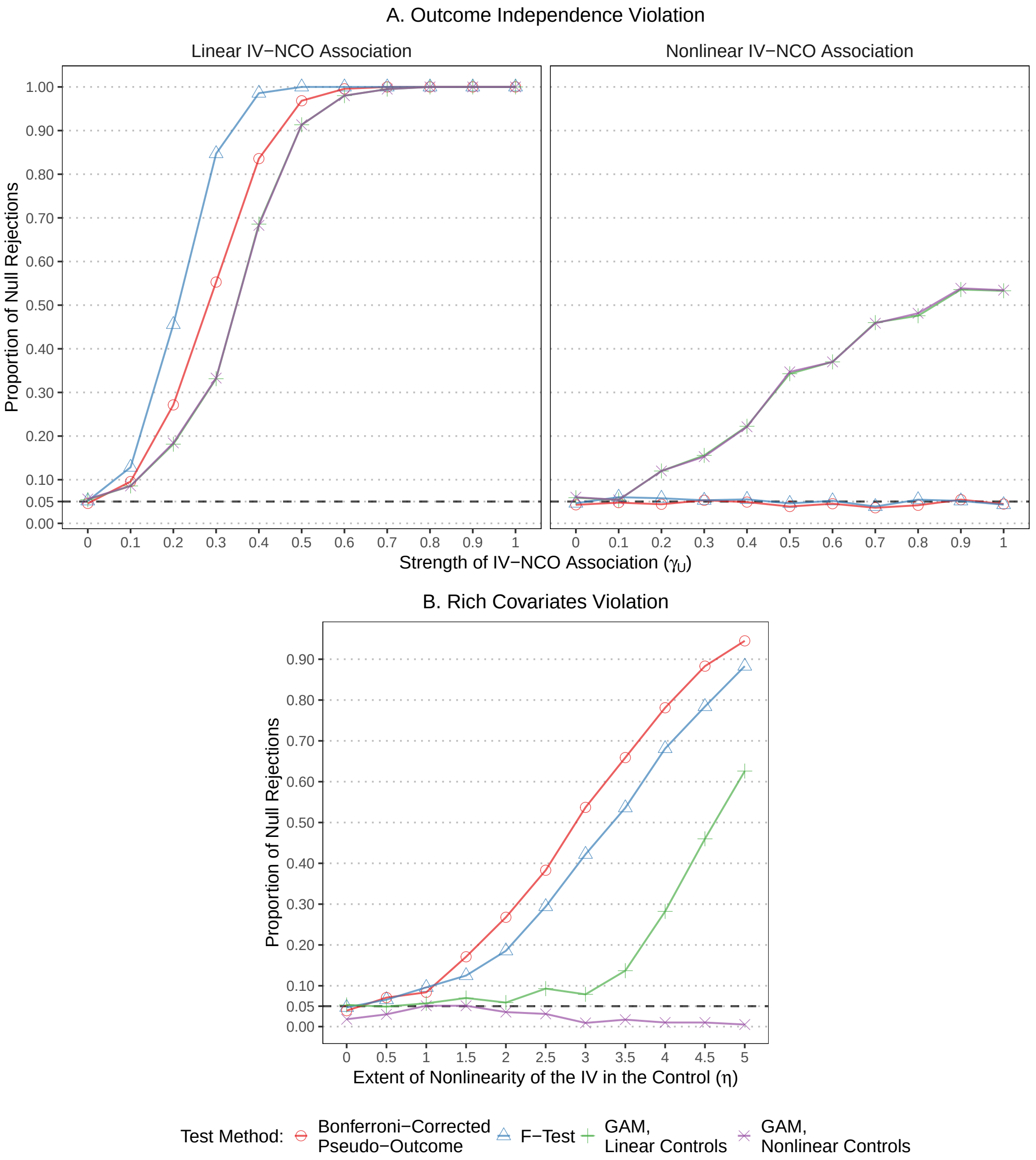


(A) An unconditional NCI test

(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 red 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, in the sense that if it is related to the outcome (through the dashed red 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 $NC \to Z \leftarrow U \to Y$ exists and hence $U \not\perp\!\!\!\perp Y|Z$.

Figure 3: Simulations of NCO Tests with Violations of Independence and Rich Covariates



Notes: This figure presents the results of simulation studies of NCO tests. Panel A (top) shows the H_0 rejection rate of different negative control tests with varying degrees of outcome independence violation. The x-axis shows the IV and APO variable relationship strength (γ_U from Appendix Equation (A8)), which also determines the IV and NCO relationship strength. In the left plot, the association between the APO variable and the IV is linear. In the right plot, it is highly nonlinear. Panel B (bottom) shows rejection rates in a scenario where the IV is exogenous but the rich covariates assumption is violated. The x-axis shows the level of nonlinearity of the IV in the controls (η from Appendix Equation (A9)). The data-generating process for each panel is detailed in Appendix E. Different NCO tests are represented by each line, including multiple pseudo-outcome regressions (Equation (6)) with Bonferroni correction, a single multivariate linear regression (Equation (8)) with an F-test, and a GAM with Wald test, both with and without smooth terms for continuous controls (Equations (9) and (10)). See Section 3 for detailed descriptions of the testing methods. Each simulation scenario comprised 1,000 sampled datasets of 10,000 observations.

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 et al. (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 et al. (2023)	Hospital assignment	Health outcomes	Ambulance company assignment, which strongly predicts hospital assignment	Patient ambulance assignment may depend on their health	Patient demographics
Kirkeboen et al. (2016)	Admission to field/institution	Log wages	Admission cutoff: RD design	Some students might be able to manipulate position relative to cutoff	Predicted wage based on predetermined covariates

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)
Acemoglu et al. (2020)	Socialist organization (Peasant Fasci)	Sicilian Mafia presence	The 1893 drought: Led to increase in support for socialist organizations	Weather conditions might generate convenient economic conditions for Mafia emergence	Rainfall in previous years
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 et al. (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	Pseudo-Outcome	Pseudo-IV	Pseudo-IV, Controlling for IV	# Negative Controls Used (median)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All	140	0.51	0.72	0.25	0.21	0.40	0.24	0.02	3.50
by Journal:									
REStud	48	0.42	0.75	0.40	0.10	0.65	0.40	0.02	4.00
AER	42	0.50	0.81	0.29	0.05	0.52	0.29	0.00	4.00
JPE	21	0.62	0.54	0.15	0.31	0.15	0.08	0.05	3.50
QJE	19	0.68	0.77	0.15	0.46	0.15	0.15	0.05	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 employing 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 to one, as some papers employed multiple test types. Columns (6)–(8) give the share of papers using pseudo-outcome designs, pseudo-IV designs, and pseudo-IV designs that also condition 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:	Pseudo-Outcome (single)	Pseudo-Outcome (single)	Pseudo-Outcome (multiple)	F-Test	GAM, Linear Controls	GAM, Nonlinear Controls	Number of Observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Autor et al. (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 et al. (2013) and Deming (2014). Column (1) replicates one of the original falsification analyses, in which Autor et al. used a pseudo-outcome and 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 pseudo-outcome test using one NCO. For Autor et al. the single pseudo-outcome 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 pseudo-outcome tests using all the NCOs. Column (4) uses an F-test (Equation (8)) with all NCOs jointly. Columns (5) and (6) use GAM tests with linear and smoothed controls, respectively (Equations (9) and (10)).

Table 4: Illustrative Applications of Negative Control Instrument Tests

	Without Conditioning on the IV	**With Conditioning on the IV**			
	Pseudo IV (single)	Pseudo IV (single)	Pseudo IV (multiple)	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 pseudo-IV 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 pseudo-IV test (Equation (7)) using the same NCI as Column (1); Column (3) applies Bonferroni correction for multiple pseudo-IV tests; Column (4) uses an F-test for all NCIs.

Online Supplementary Appendices

A Additional Figures and Table

Appendix Figure A1: Steps for Implementing Negative Control Tests

1. **Select Negative Control Variables**

    - Consider all candidate NCOs or NCIs in the data that can serve as proxies for identification threats (APO or API variables).
    - Evaluate whether each candidate satisfies a negative control assumption (based on Definition 3 for NCO or Definition 5 for NCI).
    - Note that including "weak" negative controls that are unlikely to be informative may reduce test power.
    - See Section 4.1 for examples of negative control variables.

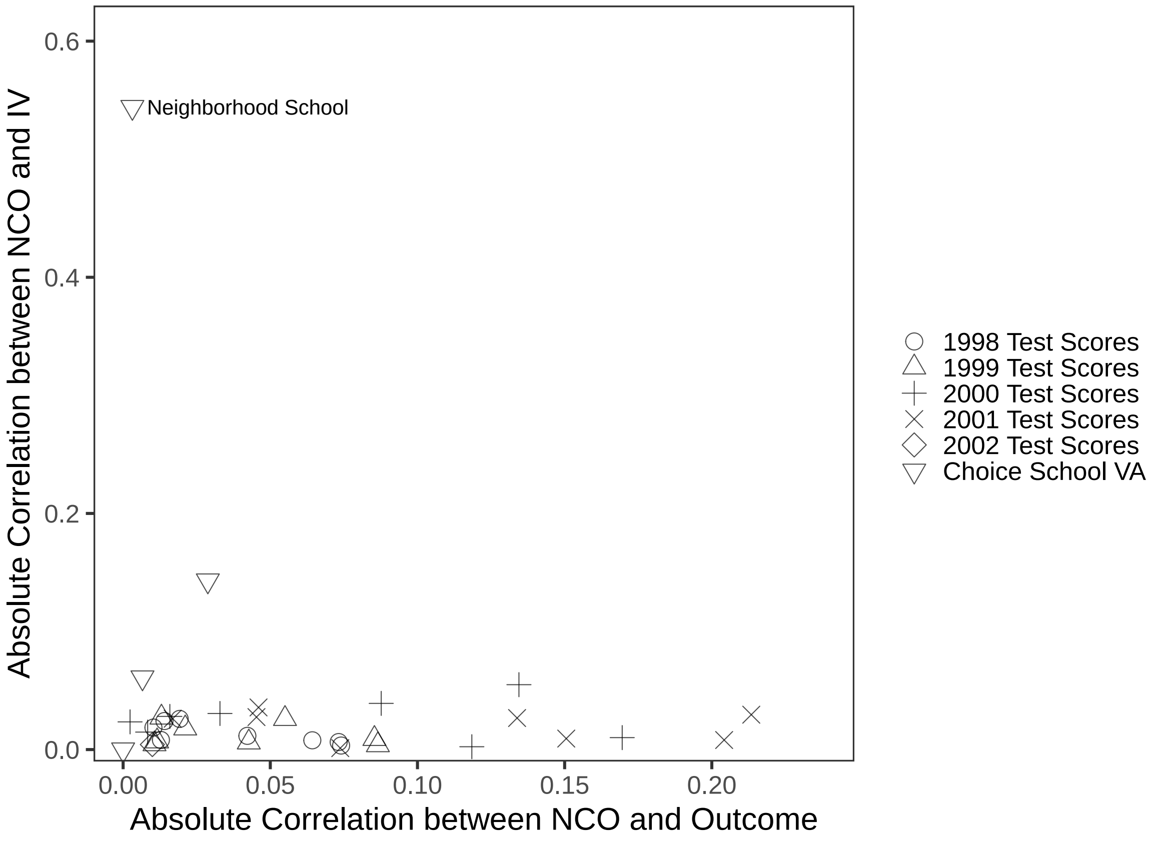
2. **Design the Negative Control Test**

    - Select the test specification: when using 2SLS and when multiple NCOs are available, consider using an F-test, which tests both IV exogeneity and rich covariates.
    - When the sample size is large, consider nonlinear methods.
    - In an NCI test, remember to condition on the IV when appropriate.
    - See Section 4.2 for discussion.

3. **Interpret Results**

    - Null rejected:
        - Test functional form and IV exogeneity separately (e.g., use Ramsey's RESET test for functional form, and a nonlinear negative control test for exogeneity).
        - Check which negative control is most related to both the outcome and the IV.
    - Null not rejected:
        - Consider statistical power.
        - Acknowledge that untested threats may still exist.
    - See Section 4.3 for discussion.

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). Correlations are calculated using data from 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 F.2 for details.

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 et al. (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., grapes)	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.

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 1.

Other Falsification Tests. As discussed in Section 1, 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 et al., 2023; Glymour et al., 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.

C Examples and Counterexamples

C.1 Non-Causal APO Variable

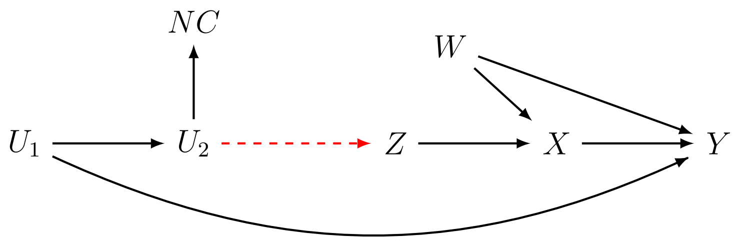


Figure C1: An illustration of causal and non-causal APO variables

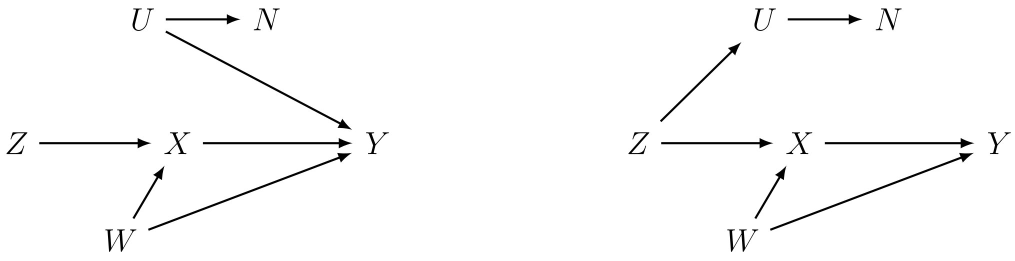
In the example presented in Figure C1, U_2 is a valid APO variable, satisfying path indication, even though it has no direct causal effect on Y (other than the possible effect through the IV). For example, assume that Z is a teacher assignment that is claimed to be quasi-random, X is the value added of the actual teacher, and Y is test scores. The variable W represents the concern that some students can switch classrooms and end up with different teachers. In this example, U_1 is unobserved ability, which directly affects test scores. U_2 represents detailed test scores in some previous exams that are also unobserved. The dashed red arrow represents a concern that principals allocate students to teachers based on the detailed test score (e.g., students with low math scores are assigned to a specific teacher).

In this example, the detailed test scores in past exams satisfy path indication even though the detailed past test scores do not directly affect future test scores. However, there is a path between the previous detailed test scores and the current test scores, because both are affected by ability (U_1).

The variable NC is aggregated previous test scores, which averages past scores in math 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 imply an alternative path from the IV to the outcome. Specifically, the presence of this path would violate outcome independence as students with different abilities would sort into different teachers based on their previous math scores.

Note that in this scenario, U_1 is also an APO variable. However, NC is a valid NCO with respect to U_2 but not with respect to U_1 alone, as, conditional on the unobserved ability, there is still a correlation between the NCO and the IV ($Z \not\perp NC|U_1$), i.e, the teacher assignment is not conditionally independent of the aggregated test scores.

C.2 Heterogeneity-Based Violation of Path Indication



(A) Females

(B) Males

Figure C2: Violation of path indication (Definition 1) due to heterogeneity

Figure C2 describes a random variable U that is associated with both the IV and the potential outcome $Y(x)$. Yet U is not an APO variable since it does not satisfy path indication. In Figure C2, the IV affects U for males but not for females. On the other hand, U affects Y for females but not for males. Path indication is not satisfied since even though the IV is exogenous ($Z \perp\!\!\!\perp Y(x)$), it is still associated with U ($Z \not\perp\!\!\!\perp U$).

Let N be a proxy for the variable U. In this example, for a large enough sample size, we would conclude that $Z \not\perp\!\!\!\perp N$ (due to the effect among males), but from this test one cannot deduce that the IV is not exogenous since U is not an APO variable.

C.3 Violation of Path Indication: Multivariate Variable

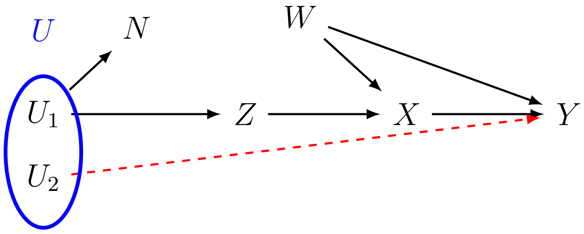


Figure C3: Violation of path indication (Definition 1) when U has multiple components

Figure C3 presents an example where U is a multivariate variable. Specifically, assume that $U = (U_1, U_2)$ is a bivariate vector of two independent variables. Assume 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 the concern that some students can switch classrooms and end up with different teachers.

Assume that U_1 is having basketball as a hobby. Assume also that basketball is correlated with the IV. For example, one teacher also coaches basketball and so basketball players (who list basketball as a hobby) are more likely to be assigned to her. However, as seen from Figure C3, a basketball hobby is independent of test scores ($U_1 \perp\!\!\!\perp Y(x)$).

Let U_2 indicate having math as a hobby. Assume that students who report math as a hobby tend to perform better in exams and that math lovers are randomly allocated across teachers ($U_2 \perp\!\!\!\perp Z$). Assume also that the basketball and math hobbies are independent ($U_1 \perp\!\!\!\perp U_2$). Finally, assume N is participation in an extracurricular basketball program (a proxy for U).

In this case, even though the vector U is associated with both the IV and the outcome, it is not an APO variable as it does not satisfy path indication. The IV is exogenous ($Z \perp\!\!\!\perp Y(x)$) even though the IV is correlated with the list of hobbies ($Z \not\perp\!\!\!\perp U$). We conclude that N is not a proper NCO. Even though $Z \perp\!\!\!\perp N|U$, it is still not an NCO since U is not an APO variable.

C.4 Multiple Threats

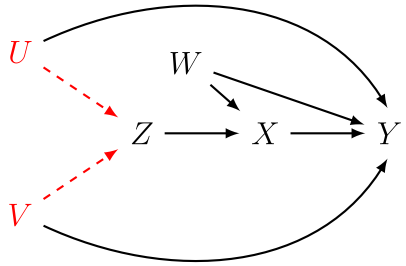


Figure C4: Multiple threats

Figure C4 presents an example of the presence of multiple threats to IV exogeneity. In this case, the variable U is an APO variable by Definition 2. 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 the concern that some students can switch classrooms and end up with different teachers. 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.

C.5 Direct IV Link Rules out Proxies of V

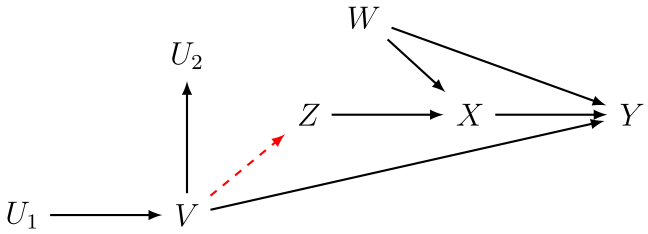


Figure C5: Violation of direct IV link (Definition 2)

Figure C5 presents the threat V (violation of outcome independence) as well as two proxies for V, U_1, and U_2. Note that latent IV exogeneity, as stated in Definition 2, 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 2, direct IV link, does not hold. If the IV is not exogenous (the dashed red 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 threat to IV exogeneity.

C.6 Path Indication Rules Out Proxies of V

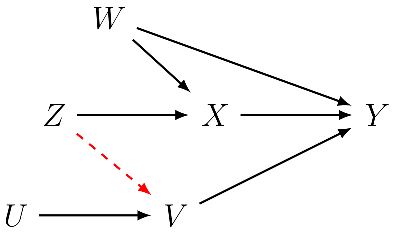


Figure C6: Violation of path indication (Definition 2)

Figure C6 presents the threat V (violation of exclusion restriction) 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 exogeneity 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 exogeneity. Condition 2 of Definition 2, namely, path indication, does not hold. Specifically, if the IV design is not exogenous (the dashed red 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.

C.7 Violation of V-validity

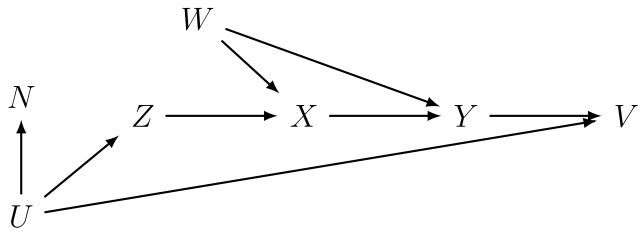


Figure C7: Violation of V-validity

Figure C7 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 2. 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 trivially 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 is exogenous ($Z \perp\!\!\!\perp Y(x)$), 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 2. Therefore U is not an APO variable, and the random variable N is not an NCO.

C.8 Violation of the NCO Assumption

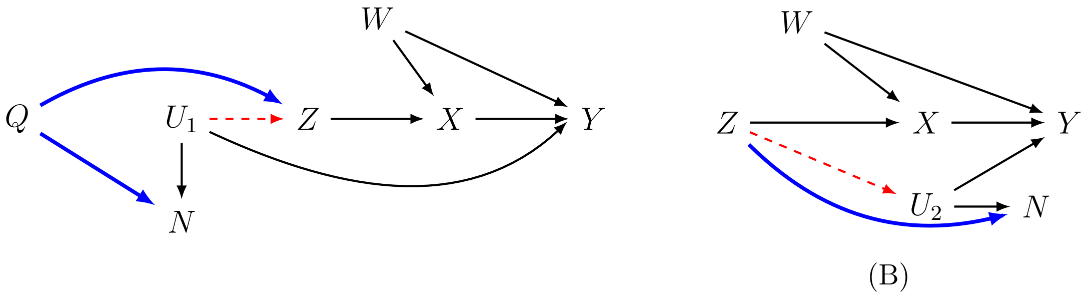
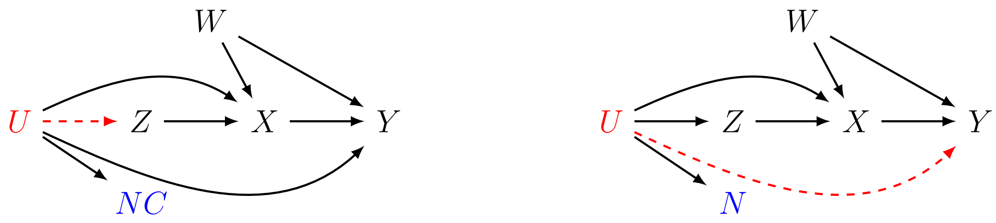


Figure C8: Violations of the Negative Control Outcome Assumption

This figure illustrates two violations of the NCO assumption. As in all other examples, 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 an IV. In these examples, U_1, U_2 are APO variables. However, N is not an NCO because it is related to Z in other ways (indicated by the solid bold blue arrows), making N uninformative about the APO variable. In Panel A, the IV and N are affected by an additional unobserved factor Q, which is unrelated to the outcome. In Panel B, the IV directly affects the variable N, violating the NCO assumption.

C.9 Potential Alternative Path Variables and Direct Treatment Link

Path indication in Definition 4 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. Figure C9 illustrates these points. In Panel A, U is a valid APO variable and the arrow $U \rightarrow X$ is allowed: latent IV exogeneity holds, and does path indication. Therefore, NC is a valid NCO, and an association between NC and Z implies that the dashed red arrow between U and Z exists, and so IV exogeneity does not hold. Conversely,



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

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

Figure C9: Direct association between potential alternative path variables and the treatment

in Panel B, U is not a valid 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 $Y \not\perp\!\!\!\perp N | Z$ does not necessarily imply violation of IV exogeneity. Intuitively, $N \not\perp\!\!\!\perp Y | Z$ even if the IV design is valid, because of the association between U and Y through X. Conditioning on X would not solve this problem, because X is a collider. Therefore, $N \not\perp\!\!\!\perp Y | Z, X$ because of the path $N \leftarrow U \rightarrow X \leftarrow W \rightarrow Y$ (Pearl, 2009).

C.10 Counterexample: A Vector of NCOs That is Not an NCO

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

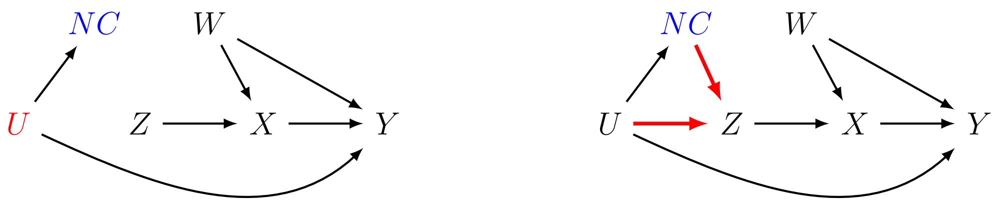
$$Z = (R_1 \oplus R_2) + \theta U + \epsilon_Z,$$

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 $Z \perp\!\!\!\perp NC_i | U$. 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 $Z \not\perp\!\!\!\perp (NC_1, NC_2) | U$ 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, $Z \not\perp\!\!\!\perp NC_2$ and NC_2 would no longer satisfy the NCO assumption.

C.11 NCO Potentially Affecting IV



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

Figure C10: Direct association between an NCO and the IV

Figure C10 presents a scenario in which if the IV is not exogenous, 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, $Z \not\perp\!\!\!\perp NC|U$, and so the lagged outcome does not satisfy Definition 3. However, when the design is exogenous, and $Z \perp\!\!\!\perp U$, the condition $Z \perp\!\!\!\perp NC|U$ is satisfied. Hence, NC is an NCO based on the broader Definition A3, defined below in Section D.1.1. Indeed, in this case, if $Z \not\perp\!\!\!\perp NC$, the design is not exogenous ($Z \not\perp\!\!\!\perp Y(x)$).

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. *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). □

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 $A \perp\!\!\!\perp Q|D$. □

D.1 Negative Controls for Instrumental Variable Designs When Control Variables Are Included

This section presents the proofs of the theoretical results from Section 2. 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 of NCO and NCI (under weaker NCO and NCI assumptions, respectively). Third, for API variables we discuss multiple threats to IV exogeneity (similar to the discussion for APO variables in Section 2.2.1).

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 $Z \perp\!\!\!\perp Y(x)|\ C$.

D.1.1 Negative Control Outcomes When Control Variables Are Included

We adapt the definitions of APO variables and NCOs as follows.

Definition A2 (Alternative path outcome variable with 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 exogeneity.* $Z \perp\!\!\!\perp Y(x)|C, U, V$.
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* $Z \perp\!\!\!\perp Y(x)|C, V$.

Definition A3 (Negative control outcome with controls)**.** *A random variable NC is an NCO if it satisfies the NCO assumption with respect to controls C: There exists an APO variable U such that if $Z \perp\!\!\!\perp U|C$ then $Z \perp\!\!\!\perp NC|\ U, C$.*

Even without controls ($C = \emptyset$), this definition is more general than Definition 3. To see this, note that in the case without controls, if $Z \not\perp\!\!\!\perp U$ (and so the design is not exogenous), Assumption A3 allows for an association $Z \not\perp\!\!\!\perp NC|\ U$. Such an association between the NCO and the IV is still informative about the validity of the IV design since this association exists only if the design is invalid. Appendix C.11 provides an example of such an NCO. Every variable that satisfies the NCO assumption in Definition 3 trivially satisfies this less restrictive definition.

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 A1. *Assume that a random variable NC satisfies the NCO assumption with respect to controls C (Definition A3). If $Z \not\perp\!\!\!\perp NC|C$ then $Z \not\perp\!\!\!\perp Y(x)|C$. That is, the IV design is not exogenous.*

Proof. We begin by showing that $Z \not\perp\!\!\!\perp NC|C$ implies that $Z \not\perp\!\!\!\perp U|C$. Else, if $Z \perp\!\!\!\perp U|C$ then by the NCO assumption (see Definition A3) $Z \perp\!\!\!\perp NC|\ U, C$. Based on Lemma 2, $Z \perp\!\!\!\perp NC|\ U, C$ and $Z \not\perp\!\!\!\perp NC|C$ 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 $Z \not\perp\!\!\!\perp Y(x)|C$. □

D.1.2 Negative Control Instruments When Control Variables Are Included

We first extend Definition 4 to allow for additional threats (similar to Definition 2 for APO variables) and to include controls C.

Definition A4 (Alternative path instrument variable with 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 exogeneity.* $Z \perp\!\!\!\perp Y(x)|C, U, V$.
2. *Path indication. If* $Z \perp\!\!\!\perp Y(x)|C, V$ *then* $U \perp\!\!\!\perp Y|Z, C, V$.
3. *Direct outcome link. If* $U \not\!\perp\!\!\!\perp Y|Z, C, V$ *then* $U \not\!\perp\!\!\!\perp Y|Z, C$.
4. *V-validity. If* $Z \not\!\perp\!\!\!\perp Y(x)|C$ *then* $Z \not\!\perp\!\!\!\perp Y(x)|C, V$.

Condition 1 is the same as in Definition A2. Conditions 2–4 together imply a version of Condition 2 from Definition 4 that includes controls. Similar to the theory of APO variables, the condition is decomposed into three independent conditions to exclude proxies for threats and maintain that V is indeed a threat (similar to Definition A2).

Next, we generalize the definition of an NCI to include controls and allow for direct associations with the outcome if the IV design is not exogenous.

Definition A5 (Negative control instrument with controls)**.** *A random variable NC is an NCI if it satisfies the NCI assumption with respect to controls C: There exists an API variable U such that if $U \perp\!\!\!\perp Y|Z, C$ then*

$$Y \perp\!\!\!\perp NC|Z, C, U. \tag{A2}$$

We are now ready to 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 A2. *Assume that a random variable NC satisfies the NCI assumption with respect to controls C (Definition A5). If $Y \not\!\perp\!\!\!\perp NC|Z, C$, then $Z \not\!\perp\!\!\!\perp Y(x)|C$. That is, the IV design is not exogenous.*

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 $Z \not\!\perp\!\!\!\perp Y(x)|C$; i.e., IV exogeneity is violated.

We now turn to the other case, where $U \perp\!\!\!\perp Y|Z, C$. By the NCI assumption (Definition A5), we have that $Y \perp\!\!\!\perp NC|Z, C, U$, and by the condition of the theorem we have that $Y \not\!\perp\!\!\!\perp NC|Z, C$. 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 A3. *Assume that a random variable NC satisfies the NCI assumption with respect to controls C (Definition A5). If, in addition, $Z \perp\!\!\!\perp NC|C$, then if $Y \not\perp\!\!\!\perp NC|C$, then $Z \not\perp\!\!\!\perp Y(x)|C$.*

Proof. Assume by way of contradiction that IV exogeneity $Z \perp\!\!\!\perp Y(x)|C$ holds. Because NC is an NCI, it follows from Theorem A2 that $Y \perp\!\!\!\perp NC|Z, C$. Additionally, based on the assumption that $Z \perp\!\!\!\perp NC|C$, Lemma 1 implies that $Y \perp\!\!\!\perp NC|C$, which contradicts the premise. □

D.2 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

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

Because outcome independence, exclusion restriction, and rich covariates are assumed, Equation (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. Similar to the proof of Corollary 1, we write the equivalent minimization problem as

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

Because outcome independence, exclusion restriction, and CSRF are assumed, Equation (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.

□

Proof of Corollary 3

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.

□

D.3 Bias Correction

In this section, we first present an alternative assumption instead of treatment effect homogeneity, under which τ^{DiW} identifies a causal effect. Then, we discuss additional scenarios for the period 0 assignments of the IV and the treatment. Finally, in Section D.3.3, we present proofs for Theorem 4 in the main text, and for the appendix theorems.

D.3.1 No Treatment without IV

Theorem 4 showed a set of assumptions under which the DiW estimator identifies a causal effect. An alternative set of assumptions can be invoked when treatment is only available for observations that receive the IV.

Assumption A3 (No treatment without IV)**.** $X_1(0) = 0$.

Using terminology from Angrist et al. (1996), this assumption implies that there are no always-takers and defiers in the population. Therefore, monotonicity ($X_1(1) \geq X_1(0)$) holds as well. For example, Assumption A3 is satisfied if the IV is a voucher for receiving treatment, and treatment is not available without a voucher.

When this assumption holds, the DiW estimator can be used for bias correction even without treatment effect homogeneity. Specifically, τ^{DiW} identifies the treatment effect for the treated compliers.

Theorem A4. *Assume that treatment independence, exclusion restriction, Assumption 5, and Assumption A3 hold. The difference-in-Wald estimator identifies the causal effect on the treated compliers. That is,*

$$\tau^{DiW} = \mathbb{E}[Y_1(1) - Y_1(0)|X_1 = 1, X_1(1) > X_1(0)].$$

The proof is given in Section D.3.3.

D.3.2 Additional Scenarios

In Section 5, we discussed the scenario of an NCO Y_0, which is a lagged outcome from a preperiod before the IV and the treatment were initiated (i.e., $Z_0 = X_0 = 0$ with probability one). In this section, we study two additional scenarios for the preperiod.

- No IV: $Z_0 = 0, X_0 = X_0(0)$.
- No IV effect on the treatment: $Z_0 = Z_1, X_0 = X_0(0)$.

In the no-IV scenario, the treatment has already started; however, the IV has not been initiated yet. This corresponds to cases where the IV represents a new policy to encourage participation in an ongoing treatment. For example, the IV may be a new subsidy for participation in an already running program.

The no IV effect on the treatment scenario considers the case where the IV has already been initiated, but the IV does not affect the treatment at the preperiod. Moreover, we assume that the IV is the same in both periods. The treatment has also already been initiated. However, the IV still does not have a causal effect on the treatment. For example, the IV could be a subsidy to encourage participation in the treatment. In contrast to the no-IV scenario, here the subsidy has already been given at the preperiod. Yet, it still does not affect the treatment (e.g., there is a delay in joining the treatment, which does not affect those already treated). While this scenario is most useful for bias correction, it is also quite seldom observed.

As before, we assume that outcome independence, relevance, and exclusion restriction hold. Since we do not have a homogeneous treatment effect, we also need to assume that monotonicity ($X_1(1) \geq X_1(0)$) holds as well. We also assume that the NCO satisfies similar assumptions to the assumptions of DiD (Roth et al., 2023). Denote by $Y_t(z_0, x_0, z_1, x_1)$ the potential outcomes at time $t = 0, 1$. For brevity we sometimes write $Y_t(z_s = a, x_s = b)$ to denote an intervention on z_s, x_s. In this case z_r, x_r for $r \neq s$ are not intervened ($z_r = Z_r, x_r = X_r$).

Assumption A4 (DiD for IV assumptions)**.**

1. No Anticipation $Y_0(z_0, x_0, z_1, x_1) = Y_0(z_0, x_0, 0, 0)$ *for every value of* $z_0, x_0, z_1, x_1 \in \{0, 1\}^4$.

2. Parallel Trends

$$\begin{aligned} &\mathbb{E}\big[Y_1(z_1 = Z_0, x_1 = X_0(Z_0)) - Y_0(z_0 = Z_0, x_0 = X_0(Z_0)) \mid Z_1 = 1\big] \\ &= \mathbb{E}\big[Y_1(z_1 = Z_0, x_1 = X_0(Z_0)) - Y_0(z_0 = Z_0, x_0 = X_0(Z_0)) \mid Z_1 = 0\big]. \end{aligned}$$

The first condition implies that the lagged outcome Y_0 is not affected by the future treatment or IV. The second condition generalizes the DiD parallel-trends assumption for the IV setting. It is a counterfactual statement that in the absence of changes in the IV and the treatment between periods, the trends in the outcome would be the same in expectation for observations that are IV-treated and IV-untreated. In the case of no IV and no treatment ($Z_0 = X_0 = 0$), this assumption boils down to Condition 2 in Assumption 5.

Since in these scenarios treatment is defined in the preperiod as well, we need to make an additional assumption.

Assumption A5 (Same type)**.** *For every value of* $z \in \{0, 1\}$, $X_0(z) = X_1(z)$.

This assumption states that for all observations, the observation type as defined by Angrist et al. (1996) (always-taker, compiler, never-taker) is the same in both periods. Hence, always-takers are treated in both periods, never-takers are untreated in both periods, and compliers are only treated when the IV equals one in period 1.

The following two theorems establish the identification of causal effects by τ^{DiW} in each of the two scenarios.

Theorem A5 (No IV)**.** *Assume the no-IV scenario. Under treatment independence, relevance, monotonicity, exclusion restriction, Assumption A4, and same type (Assumption A5), the difference-in-Wald estimator identifies the causal effect of the treatment on the treated compliers. That is,*

$$\tau^{DiW} = \mathbb{E}[Y_1(1) - Y_1(0) | X_1 = 1, X_1(1) > X_1(0)].$$

Theorem A6 (No IV effect on the treatment)**.** *Assume the no IV effect on the treatment scenario. Under treatment independence, relevance, monotonicity, Assumption A4, and same type (Assumption A5), the difference-in-Wald estimator identifies the causal effect on the treated compliers. That is,*

$$\tau^{DiW} = [Y_1(1,1) - Y_1(1,0)|X_1 = 1, X_1(1) > X_1(0)].$$

Theorem A5 shows that in the no-IV scenario, *DiW* can be used for bias correction in case of a violation of outcome independence. However, a violation of the exclusion restriction would still bias this estimator. Theorem A6 shows that in the no IV effect on the treatment scenario, *DiW* could address violations of both outcome independence and exclusion restriction. In both cases, the *DiW* estimator estimates the causal effect only for the treated compliers. Note that in Theorem A6 we have not assumed the exclusion restriction, hence for clarity the value of the IV appears in the definition of the effect.

D.3.3 Proofs

We first prove the following lemma that will be used for the proofs of Theorems 4 and A4.

Lemma 3. *Assume $Z_0 = X_0 = 0$ and assume that treatment independence, relevance, exclusion restriction, and Assumption 5 hold. The difference-in-Wald estimator equals*

$$\tau^{DiW} = \frac{\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(1) > X_1(0)\big] \Pr\big[X_1(1) > X_1(0)\big] + B_1}{\mathbb{E}[X_1|Z_1 = 1] - \mathbb{E}[X_1|Z_1 = 0]}, \tag{A3}$$

where

$$\begin{aligned} B_1 = &\Pr\Big(X_1(0) = 1, X_1(1) = 1\Big)\Big(\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(0) = 1, X_1(1) = 1\big] \\ &- \mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 0, X_1(0) = 1, X_1(1) = 1]\Big) \\ &- \Pr\Big(X_1(0) = 1, X_1(1) = 0\Big)\Big(\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 0, X_1(1) < X_1(0)]\Big). \end{aligned}$$

Proof. By the no-anticipation condition, $Y_0(z_1, x_1) = Y_0(0,0)$, and hence we can write $Y_0 = Y_0(0,0)$. The numerator of the DiW estimator then equals

$$\begin{aligned} &\mathbb{E}[Y_1 - Y_0|Z_1 = 1] - \mathbb{E}[Y_1 - Y_0|Z_1 = 0] \\ &\qquad = \mathbb{E}\big[Y_1(1, X_1(1)) - Y_0(0,0)|Z_1 = 1\big] - \mathbb{E}\big[Y_1(0, X_1(0)) - Y_0(0,0)|Z_1 = 0\big]. \end{aligned} \tag{A4}$$

By the parallel-trends condition, we can replace $Y_0(0,0)$ with $Y_1(0,0)$ inside both of the expectations of (A4) to obtain

$$\mathbb{E}[Y_1(1, X(1)) - Y_1(0,0)|Z_1 = 1] - \mathbb{E}[Y_1(0, X(0)) - Y_1(0,0)|Z_1 = 0]. \tag{A5}$$

By the law of total expectation with respect to the type (compliers, for which $X_1(1) > X_1(0)$; never-takers, for which $X_1(1) = X_1(0) = 0$; always-takers, for which $X_1(1) = X_1(0) = 1$; and defiers, for which $X_1(1) < X_1(0)$), and by treatment independence $\{X(0), X(1)\} \perp\!\!\!\perp Z$, we can write Equation (A5) as

$$\begin{aligned}
&\Pr\big[X_1(1) > X_1(0)\big]\Big(\mathbb{E}\big[Y_1(1,1) - Y_1(0,0)|Z_1 = 1, X_1(1) > X_1(0)\big] - \mathbb{E}\big[Y_1(0,0) - Y_1(0,0)|Z_1 = 0, X_1(1) > X_1(0)\big]\Big)\\
&+ \Pr\big(X_1(1) = X_1(0) = 0\big)\Big(\mathbb{E}\big[Y_1(1,0) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = X_1(0) = 0\big] - \mathbb{E}\big[Y_1(0,0) - Y_1(0,0) \mid Z_1 = 0, X_1(1) = X_1(0) = 0\big]\Big)\\
&+ \Pr\big(X_1(1) = X_1(0) = 1\big)\Big(\mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = X_1(0) = 1\big] - \mathbb{E}[Y_1(0,1)] - Y_1(0,0) \mid Z_1 = 0, X_1(1) = X_1(0) = 1\big]\Big)\\
&+ \Pr\big[X_1(1) < X_1(0)\big]\Big(\mathbb{E}\big[Y_1(1,0) - Y_1(0,0)|Z_1 = 1, X_1(1) < X_1(0)\big] - \mathbb{E}\big[Y_1(0,1) - Y_1(0,0)|Z_1 = 0, X_1(1) < X_1(0)\big]\Big)\\
&= \Pr\big[X_1(1) > X_1(0)\big]\mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid Z_1 = 1, X_1(1) > X_1(0)\big]\\
&+ \Pr\big(X_1(1) = X_1(0) = 0\big)\mathbb{E}\big[Y_1(1,0) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = X_1(0) = 0\big]\\
&+ \Pr\big[X_1(1) = X_1(0) = 1\big]\Big(\mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = X_1(0) = 1\big] - \mathbb{E}\big[Y_1(0,1) - Y_1(0,0) \;|Z_1 = 0, X_1(1) = X_1(0) = 1\big]\Big)\\
&+ \Pr\big[X_1(1) < X_1(0)\big]\Big(\mathbb{E}\big[Y_1(1,0) - Y_1(0,0)|Z_1 = 1, X_1(1) < X_1(0)\big] - \mathbb{E}\big[Y_1(0,1) - Y_1(0,0)|Z_1 = 0, X_1(1) < X_1(0)\big]\Big).
\end{aligned}$$

By the exclusion restriction assumption this becomes

$$\begin{aligned}
&\Pr\big[X_1(1) > X_1(0)\big]\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(1) > X_1(0)\big]\\
&+ \Pr\big[X_1(1) = X_1(0) = 1\big]\Big(\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(1) = X_1(0) = 1\big] - E\big[Y_1(1) - Y_1(0) \;|Z_1 = 0, X_1(1) = X_1(0) = 1\big]\Big)\\
&- \Pr\big[X_1(1) < X_1(0)\big]\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 0, X_1(1) < X_1(0)\big]
\end{aligned}$$

The denominator equals $\mathbb{E}[X_1|Z_1 = 1] - \mathbb{E}[X_1|Z_1 = 0]$ and hence we have the expression of Equation (A3). □

Proof of Theorem 4

Proof. By Lemma 3 and Assumption 6, the numerator of τ^{DiW} is

$$\Pr\big[X_1(1) > X_1(0)\big]\tau - \Pr\big[X_1(1) < X_1(0)\big]\tau.$$

The denominator of τ^{DiW} (the first stage) identifies

$$\Pr\big[X_1(1) = 1\big] - \Pr\big[X_1(0) = 1\big] = \Pr\big[X_1(1) > X_1(0)\big] - \Pr\big[X_1(1) < X_1(0)\big].$$

Dividing the numerator by the denominator yields $\tau^{DiW} = \tau$.

□

Proof of Theorem A4

Proof. By Lemma 3, the numerator of τ^{DiW} equals

$$\mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(1) > X_1(0)\big] \Pr\big[X_1(1) > X_1(0)\big] + B_1.$$

Since there are no always-takers or defiers (by Assumption A3), we have that $\Pr(X_1(0) = 1) = 0$ and therefore $B_1 = 0$. Furthermore, the denominator (the first stage) identifies $\Pr\big[X_1(1) > X_1(0)\big]$. Therefore,

$$\tau^{DiW} = \mathbb{E}\big[Y_1(1) - Y_1(0) \mid Z_1 = 1, X_1(1) > X_1(0)\big].$$

□

Proof of Theorem A5

Proof. Under the no-IV scenario, $Z_0 = 0, X_0 = X_0(0)$, and the parallel-trends condition takes the form

$$\begin{aligned} &\mathbb{E}\big[Y_1(z_1 = 0, x_1 = X_0(0)) - Y_0(z_0 = 0, x_0 = X_0(0)) \mid Z_1 = 1\big] \\ &= \mathbb{E}\big[Y_1(z_1 = 0, x_1 = X_0(0)) - Y_0(z_0 = 0, x_0 = X_0(0)) \mid Z_1 = 0\big]. \end{aligned}$$

The numerator of the DiW estimator is

$$\begin{aligned} &\mathbb{E}\big[Y_1 - Y_0 \mid Z_1 = 1\big] - \mathbb{E}\big[Y_1 - Y_0 \mid Z_1 = 0\big] \\ &\qquad = \mathbb{E}\Big[Y_1\big(z_1 = 1, x_1 = X_1(1)\big) - Y_0\big(z_0 = 0, x_0 = X_0(0)\big) \mid Z_1 = 1\Big] \\ &\qquad - \mathbb{E}\Big[Y_1\big(z_1 = 0, x_1 = X_1(0)\big) - Y_0\big(z_0 = 0, x_0 = X_0(0)\big) \mid Z_1 = 0\Big] \\ &\qquad = \mathbb{E}\Big[Y_1\big(z_1 = 1, x_1 = X_1(1)\big) - Y_1\big(z_1 = 0, x_1 = X_0(0)\big) \mid Z_1 = 1\Big] \\ &\qquad - \mathbb{E}\Big[Y_1\big(z_1 = 0, x_1 = X_1(0)\big) - Y_1\big(z_1 = 0, x_1 = X_0(0)\big) \mid Z_1 = 0\Big] \\ &\qquad = \mathbb{E}\Big[Y_1\big(z_1 = 1, x_1 = X_1(1)\big) - Y_1\big(z_1 = 0, x_1 = X_1(0)\big) \mid Z_1 = 1\Big], \end{aligned}$$

where the first equality is by definition, the second equality by the parallel-trends condition and the third equality follows from the fact that under same type $X_0(0) = X_1(0)$.

As in the proof of Lemma 3, we continue by the law of total expectation with respect to the type and use treatment independence to obtain

$$\begin{aligned}
&\mathbb{E}\big[Y_1(1, X_1(1)) - Y_1(0, X_1(0)) \mid Z_1 = 1\big] \\
&= \Pr[X_1(1) > X_1(0)]\mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid Z_1 = 1, X_1(1) > X_1(0)\big] \\
&+ \Pr\big[X_1(1) = 0\big]\mathbb{E}\big[Y_1(1,0) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = 0\big] \\
&+ \Pr\big[X_1(0) = 1\big]\mathbb{E}\big[Y_1(1,1) - Y_1(0,1) \mid Z_1 = 1, X_1(0) = 1\big].
\end{aligned}$$

Because of treatment independence and monotonicity, the denominator equals $\Pr[X_1(1) > X_1(0)]$. Dividing the numerator by this denominator, we have that

$$\begin{aligned}
DiW &= \mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid Z_1 = 1, X_1(1) > X_1(0)\big] + B_2 \\
&= \mathbb{E}\big[Y_1(1,1) - Y_1(0,0) \mid X_1 = 1, X_1(1) > X_1(0)\big] + B_2,
\end{aligned} \tag{A6}$$

where

$$\begin{aligned}
B_2 = &\frac{\Pr(X_1(1) = 0)}{\Pr[X_1(1) > X_1(0)]}\mathbb{E}[Y_1(1,0) - Y_1(0,0) \mid Z_1 = 1, X_1(1) = 0] \\
+ &\frac{\Pr(X_1(0) = 1)}{\Pr[X_1(1) > X_1(0)]}\mathbb{E}\big[Y_1(1,1) - Y_1(0,1) \mid Z_1 = 1, X_1(0) = 1\big].
\end{aligned} \tag{A7}$$

By the exclusion restriction assumption, the first term in (A6) becomes

$$\mathbb{E}\big[Y_1(1) - Y_1(0) \mid X_1 = 1, X_1(1) > X_1(0)\big]$$

and, furthermore, $B_2 = 0$ because $Y_1(1,0) = Y_1(0,0)$ and $Y_1(1,1) = Y_1(0,1)$. □

Proof of Theorem A6

Proof. By the no IV effect on the treatment, $Z_0 = Z_1, X_0 = X_0(0)$, so the parallel-trends condition takes the form

$$\begin{aligned}
&\mathbb{E}\big[Y_1(z_1 = 1, x_1 = X_0(0)) - Y_0(z_0 = 1, x_0 = X_0(0)) \mid Z_1 = 1\big] \\
&= \mathbb{E}\big[Y_1(z_1 = 0, x_1 = X_0(0)) - Y_0(z_0 = 0, x_0 = X_0(0)) \mid Z_1 = 0\big].
\end{aligned}$$

As before, we start with the numerator:

$$\begin{aligned}
&\mathbb{E}\big[Y_1 - Y_0 \mid Z_1 = 1\big] - \mathbb{E}[Y_1 - Y_0 \mid Z_1 = 0\big] \\
&\qquad = \mathbb{E}\big[Y_1(z_1 = 1, x_1 = X_1(1)) - Y_0(z_0 = 1, x_0 = X_0(0)) \mid Z_1 = 1\big] \\
&\qquad - \mathbb{E}\big[Y_1(z_1 = 0, x_1 = X_1(0)) - Y_0(z_0 = 0, x_0 = X_0(0)) \mid Z_1 = 0\big] \\
&\qquad = \mathbb{E}\big[Y_1(z_1 = 1, x_1 = X_1(1)) - Y_1(z_0 = 1, x_0 = X_0(0)) \mid Z_1 = 1\big] \\
&\qquad - \mathbb{E}\big[Y_1(z_1 = 0, x_1 = X_1(0)) - Y_1(z_1 = 0, x_1 = X_0(0)) \mid Z_1 = 0\big] \\
&\qquad = \mathbb{E}\big[Y_1(z_1 = 1, x_1 = X_1(1)) - Y_1(z_1 = 0, x_1 = X_1(0)) \mid Z_1 = 1\big],
\end{aligned}$$

where the first equality is by definition, the second equality by the parallel trends condition, and the third equality follows the same type assumption, $X_0(0) = X_1(0)$.

Now, by the law of total expectation the last expression equals

$$\Pr\big[X_1(1) > X_1(0)\big]\mathbb{E}\big[Y_1(1, X_1(1)) - Y_1(1, X_0(0)) \mid Z_1 = 1, X_1(1) > X_1(0)\big],$$

because for the always-takers and for the never-takers, $Y_1(1, X_1(1)) = Y_1(1, X_1(0))$.

Because of treatment independence and monotonicity, the denominator (the first stage) identifies $\Pr[X_1(1) > X_1(0)]$. Dividing the numerator by the first stage yields

$$\mathbb{E}\big[Y_1(1, X_1(1)) - Y_1(1, X_0(0)) \mid Z_1 = 1, X_1(1) > X_1(0)\big].$$

□

E Simulation Details

This section provides additional details about the DGPs used for the simulations in Section 3.5. In these analyses, we specify the relationships between the IV, the controls, the alternative path variables, and the negative controls. We do not specify the DGP of the outcome and exactly how the APO variable is associated with it, as the NCO test performance does not depend on this information. For all scenarios described in this section, and for each unique combination of parameter values, we simulated 1,000 datasets, each with 10,000 observations.

Violations of Outcome Independence. For the first analysis, the DGPs were parameterized as follows.

$$\begin{aligned} Z &= \gamma_0 + \gamma_U g(U) + \sum_{j=1}^{5} C_j + \epsilon_Z, \\ NC_k &= \beta U + \epsilon_{NC,k}, \end{aligned} \tag{A8}$$

where Z is the IV, U is an APO variable, C is a vector of five independent control variables indexed by j, NC is a vector of ten NCOs indexed by k, and $\epsilon_Z, \epsilon_{NC,k}$ are independent normally distributed error terms.

The APO variable was simulated from a uniform distribution $U \sim U[-3, 3]$. The vector of the five independent control variables C included $C_1, C_2, C_3 \sim N(1, 3)$, $C_4 \sim Ber(0.3)$, and $C_5 \sim Ber(0.5)$. The intercept was taken to be $\gamma_0 = 1$. For γ_U (the parameter controlling the magnitude of the APO variable association with the IV) we took the values $\gamma_U = 0, 0.1, 0.2, ..., 1$. The error term for Z was simulated from $\epsilon_Z \sim \mathcal{N}(0, 5)$, and for each k, $\epsilon_{NC,k}$ was simulated independently as $\epsilon_{NC,k} \sim \mathcal{N}(0, \sigma^2)$.

We consider two alternative scenarios. In the first, "linear" scenario (Panel A of Figure 3), $g(u) = u$ (the identity) and $\sigma^2 = 25$. In the second, "nonlinear" scenario (Panel B), $g(u) = 8 \min\{u^2, 1.5\}$ (a truncated parabolic function) and $\sigma^2 = 3$.

Violations of Rich Covariates. We use the same DGP and parameter values specified by Blandhol et al. (2022) in their simulation study. Following Blandhol et al. (2022), the IV Z is a binary variable with a probability that can be written as a cubic polynomial in a single control variable, drawn from a uniform distribution $C \sim U(0, 1)$. That is, $\Pr(Z = 1|C) = \gamma_0 + \gamma_1 C - \gamma_2 C^2 + \gamma_3 C_i^3$, with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.119, 1.785, -1.534, 0.597)$.

We augment this DGP with an APO variable U (simulated from a uniform distribution $U \sim U[-3, 3]$ as before). We then define ten negative controls associated with this APO. To that end, each NCO is simulated by

$$NC = \beta_1 C + \eta(\beta_0 + \beta_2 C^2 + \beta_3 C^3) + \beta_U U + \epsilon_{NC}, \tag{A9}$$

where ϵ_{NC} is simulated from a standard normal distribution $\epsilon_{NC} \sim N(0, 1)$. We take $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.119, 1.785, -1.534, 0.597)$ as well and $\beta_U = 0.3$.

In this DGP, the parameter η is a weight on the nonlinear part of the association between C and NC. As η increases the association becomes more nonlinear. For this parameter, we take the values $\eta = 0.0.5, 1, 1.5, ..., 5$.

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

This section provides additional details for the analysis in Section 4, 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.

F.1 Autor, Dorn and Hanson (2013)

Sample Construction. For this analysis, we use the original study's data from Autor et al. (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 occurred for reasons unrelated to Chinese import. 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 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 rest 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 used originally 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), 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 the future treatments (years 1990 and 2000 averaged), instrumented by the future IVs (see Column (4) of Table 2 in ADH). We report the p-value of the coefficient on the treatment with cluster robust (by state) standard errors. 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).

F.2 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 can as

$$IV_i = L_i VAM_i^1 + (1 - L_i)\, VAM_i^N \tag{A10}$$

[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.

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 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 F-test and pseudo-outcome with Bonforoeni correction, we perform a fixed-effect regression 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. In the pseduo-outcome analysis with a single NCO, we repeat the analysis with lagged test scores (from 2001–2002) and a linear specification.

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.

F.3 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 exogeneity separately.

F.4 Ashraf and Galor (2013)

Sample Construction. The study's 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 distance 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.