INSTITUTE for REPLICATION

No. 1
I4R DISCUSSION PAPER SERIES

Spurious Regressions and Panel IV Estimation: Revisiting the Causes of Conflict

Paul Christian

Christopher B. Barrett

September 2022

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 1

Spurious Regressions and Panel IV Estimation: Revisiting the Causes of Conflict

Paul Christian*, Christopher B. Barrett**

* *DIME, World Bank*
** *H. Dyson School of Applied Economics and Management, Jeb E. Brooks School of Public Policy, Cornell University*

SEPTEMBER 2022

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/ Germany
www.i4replication.org

ISSN: 2752-1931

Spurious Regressions and Panel IV Estimation: Revisiting the Causes of Conflict

PAUL CHRISTIAN AND CHRISTOPHER B. BARRETT*

Abstract: The long-recognized spurious regressions problem can lead to mistaken inference in panel instrumental variables (IV) estimation. Spurious correlations arising from correlated cycles in finite time horizons can make irrelevant instruments appear strong with signable consequences for estimated IV coefficients, or interfere with valid of inference of causal effects from IV coefficients estimated using relevant instruments. The inclusion of time fixed effects in interacted specifications does not always resolve these problems. We demonstrate these concerns by revisiting recent studies of the causal origins of conflict. We offer diagnostic and corrective recommendations for avoiding the pitfalls arising from time series exhibiting persistence.

Keywords: *Instrumental Variables, Conflict, Economic Shocks, Panel Data*

* Christian: DIME, World Bank (1818 H St, Washington DC 20433, email: pchristian@worldbank.org). Barrett: Charles H. Dyson School of Applied Economics and Management and Jeb E. Brooks School of Public Policy, Cornell University (340D Warren Hall, Ithaca, NY, 14853, email: cbb2@cornell.edu). An earlier version circulated with the title "Revisiting the Effect of Food Aid on Conflict: A Methodological Caution." Thank you to Jenny Aker, Marc Bellemare, Aureo de Paula, Brian Dillon, Teevrat Garg, Bruce Hansen, Rema Hanna, Sylvan Herskowitz, Peter Hull, Masumai Imai, David Jaeger, Joe Kaboski, Eeshani Kandpal, John Leahy, Erin Lentz, Shanjun Li, Stephanie Mercier, Francesca Molinari, Nathan Nunn, Debraj Ray, Steven Ryan, Hans-Joachim Voth, two anonymous reviewers, and seminar audiences at Cornell, Minnesota, Notre Dame, Otago, Tufts, UC-Davis, Waikato, the World Bank, NEUDC, and the Midwest International Economic Development Conference for helpful comments, and to Utsav Manjeer for excellent research assistance. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

A substantial empirical literature explores important questions that do not lend themselves to experimentation – like the causes of violent conflict – using panel data instrumental variables (IV) estimation methods to achieve causal identification using a plausibly exogenous time series variable as the IV. In this paper, we show that seemingly unobjectionable, exogenous time series instruments pose a problem for causal inference when one fails to address the time series process underlying the instrument and the dependent and explanatory variables of interest. The central issue is that one cannot ignore the sequencing of observations in the panel.

It has long been known that conventional tests that assume constant error variance over time over-reject the zero-impact null hypothesis in the presence of variables that exhibit a common time series trend, a phenomenon known in single time series as the "spurious regression" or "nonsense correlation" critique (Yule 1926, Slutsky 1937, Granger and Newbold 1973, Phillips and Hansen 1990). This was extended to the panel data difference-in-differences context by Phillips and Hansen (1990) and Bertrand et al (2004).[1]

In this paper, we explain and demonstrate how in the panel IV context spurious regressions not only lead to mistaken inference (i.e., incorrect standard errors), but also generate coefficient estimates that often have wrong magnitude or even incorrect sign relative to the true causal parameter of interest. Serial persistence in the time series dimension of panels can generate cyclicality which, if unaddressed, yields greater risk of co-movement among variables than is accounted for by conventional significance testing, thereby increasing the risk of mistaken inference.

The applied literature has largely not recognized that spurious correlations pass, via policy endogeneity or simultaneity, to both the first stage and reduced form equations in a correlated manner. Intuitively, if the time series component of the instrument coincidentally co-trends with the time series component of the outcome within a fixed time frame, the time series component of the instrument will also be coincidentally

[1] Relatedly, Nickell (1981) shows bias arises in dynamic panel data models that included lagged regressors in short panels, a different issue than we address.

correlated with the endogenous variable of interest over the same period, because the outcome and endogenous variable co-move over time on account of the endogeneity the IV is intended to solve. When instruments are strong and endogenous trends explain a small share of variation, a data generating processes with persistence will return IV coefficients further from the true causal parameters than would be expected if the time series variation is less persistent. If the variation in endogenous trends becomes large enough relative to the true causal reduced form and first stage parameters, spurious correlations can even reverse the sign of the estimated IV coefficient relative to the true causal value. In the extreme case, when endogeneity is strong and instruments are irrelevant but persistent, spurious correlation introduces a risk of the weak instruments appearing strong and statistically significant. Under plausible models, the IV coefficients estimated on these irrelevant, but incorrectly accepted instruments will always return a coefficient estimate whose sign and magnitude is determined by endogeneity rather than the causal effect of interest.

In addition to flagging the ongoing relevance of an old spurious regressions literature that has been largely overlooked in recent panel IV estimation, our approach contributes to a developing literature on the role of smooth distributions of instruments. One strand of this literature shows that discrete changes in instrumental variables can pose problems for inference. Young (2018) shows how highly leveraged observations (or clusters of observations) can bias downwards estimated standard errors in IV estimation in non-iid error processes. In our setting, the identification problem arises because inference ignores the dynamics of the instrument and other variables of interest. Because this is a mirror image of the problem studied by Young, arising from variables that are too smooth rather than too discontinuous in limited cases, we show that panel IV estimates that pass Young's test can still fall prey to spurious regressions. The problem we diagnose is similar to other cases where inference that mishandles the smoothness of the distribution of errors leads to mistaken inference. Kelly (2019) shows that failure to diagnose and account for spatial autocorrelation poses a similar problem to the spurious correlation in time series processes that we highlight.

While the inference issues with time series in panel IV are not new, we further show that one of the main fixes researchers use does not address the issue. A common panel IV estimation strategy constructs instruments by interacting a time series variable with a cross-sectional exposure variable, a special case of what are popularly known as 'shift-share' or 'Bartik (1991)' instruments. The interacted specification allows for unobserved parallel trends, bolstering the argument that identification is not affected by misspecification of omitted trends. For example, in the two examples we use to illustrate these issues, the authors argue that they identify the causal effect of inter-annual variation in the time series variable on the outcome of interest by comparing units relatively exposed to exogenous shocks to the time series instrument against relatively unexposed units. We demonstrate that the inference and bias concerns of the uninteracted specification still hold in the interacted case. When there are not multiple, independent sources of time series shocks, identification relies on a parallel trends assumption that is sometimes stated, but usually not scrutinized, and often is not satisfied in the data. When the influence of cyclicality on the outcome of interest is not constant within cross-sections of the panel, interacting the instrument with an endogenous variable and including time fixed effects or flexible trends will not solve the problem.

Our paper thus complements recent critiques of shift-share panel IV estimation. [2] We clarify that identification assumptions arising either through exogeneity of cross-sectional shift variables, as in Goldsmith-Pinkham et al (2020), or through exogeneity of time series shock variables, as in Borusyak et al. (2018), are onerous in the case of a single shock variable or highly correlated shocks and show how serial correlation creates inference and finite sample bias issues even when the instrument and outcome are independent. Our findings are perhaps closest to those of Jaeger et al. (2018) and Adão et al. (2019). Jaeger et al. (2018), studying the labor market effects of immigration, show

[2] The NQ specification is a special case of the typical Bartik setup, in which time series for multiple industries are interacted with multiple locations. In the NQ set-up, the single time series of US wheat production is analogous to having a single industry in the Bartik framework. In HI, interest rates vary across countries, but because the standard no-arbitrage condition leads to high cointegration of interest rate time series, we show that one would find identical conclusions using only the average interest rate across countries. The conclusions of this paper are therefore most relevant when exogenous variation arises mainly on one panel dimension (time or cross-section) and is either fixed or constant on the other.

that serial correlation in the treatment variable of interest biases standard shift-share instruments, but can be corrected by including a lagged endogenous regressor instrumented with a lagged shift share. Our results show that persistence in other key variables, especially the instrument, can also create finite sample bias that requires dynamic corrections. Adão et al. (2019) show via placebo simulations that conventional inference overstates rejection rates of null hypotheses in shift-share designs. We likewise use simulations to show that these issues are more general, arising from the time series properties of the instrument, dependent variable, and endogenous regressor, that they exist with or without the shift-share design, and that they arise from the spurious correlation of the distinct time series, generating biased estimates as well as inference problems. Thus, our critique extends well beyond the shift-share designs that have attracted much recent attention to unrecognized issues concerning panel IV estimation methods, in the conflict literature and beyond.

1. Revisiting the Causes of Conflict

An important thread of quantitative social science strives to identify statistically the causes of violent conflict.[3] Clean identification of causal mechanisms, even of just reduced form relationships, nonetheless remains challenging. For example, a recent systematic review focusing just on the relationship between development aid and violence identified 9,413 relevant studies, of which only 19 offered even a plausible causal identification strategy, most exploiting spatial discontinuities in within-country data from a single country (Zürcher 2017). Only five of the reviewed studies address conflict in multiple countries over time using panel data, making plausible the external validity of the findings. The most compelling cross-country studies, such as Nunn and Qian (2014, hereafter NQ), use a panel IV strategy to address the likely endogeneity of the hypothesized causal variable, in NQ's case United States (US) food aid shipments. Other recent studies of conflict use similar panel IV methods to analyze non-aid prospective causes of conflict in multi-country data. A prominent example is Hull & Imai (2015, hereafter HI), who explore

[3] Blattman and Miguel (2010) and Ray and Esteban (2017) offer excellent, accessible summaries.

the impact of gross domestic product (GDP) growth on conflict. Both HI and NQ rely on a plausibly exogenous time series instrumental variable to achieve causal identification.[4] The problem with papers that use panel IV estimation strategies similar to NQ and HI is that they implicitly ignore the sequencing of observations in the panel.

To grasp the basic intuition behind the concerns we raise, consider two descriptions of the results reported by NQ.

1. Conflict is more frequent in countries to which the US more often sends food aid following years when the US produces more wheat, relative both to countries to which the US does not frequently send food aid and to years in which the US produces less wheat.
2. An era of frequent conflict in the developing world coincided with an era of higher US wheat production. During this period, the US increased shipments of food aid and sent relatively larger shipments to those countries experiencing conflict.

While both of these statements accurately characterize NQ's findings, subtle differences in the descriptions of the dynamics lead to different inferences. The first description, which closely follows the language NQ employ, implicitly presents every year as a new experiment. One might worry that non-random aid targeting might be endogenous to conflict risks. But with enough years in the sample, it becomes "hard to think of" why conflict would be high only following years when the US experienced an exogenous, positive wheat productivity shock unless aid – which is demonstrably correlated with donor country supply shocks – causes conflict.

[4] One could choose any of a host of panel IV papers vulnerable to the spurious regressions issues we raise, on conflict and many other applications. We focus on the NQ and HI papers for a few reasons. First, the authors are exceptionally talented economists publishing in top journals; their papers represent some of the best current empirical research in the field. This underscores that the problem we address has gone largely unnoticed even among the discipline's best researchers and most rigorous peer review processes. Second, two papers is the minimum needed to establish a pattern, not a result specific to a particular paper. Third, the papers represent different forms of the broader issue we address. The endogenous regressors in each paper follow a different time series pattern, showing that this issue is not unique to a specific cyclical pattern. This paper offers a caution and some practical guidance to those pursuing panel IV estimation, not a critique of specific papers or authors.

By contrast, the second description emphasizes the time series feature, as reflected in distinct "eras". The likelihood of a coincidental alignment of US wheat production and developing country conflict depends crucially on how slowly these variables evolve and the length of the panel relative to this pace of convergence. Even a panel with 36 years may only reflect one or two distinct eras if "era" is defined as a cycle in which a single variable like US wheat production or developing country conflict is rising or falling. With slow cycles, which arise naturally from persistence in time series variables, trends can easily align by chance, spuriously, and when they do, differential correlation among sub-samples may reveal nothing if that heterogeneity results from reverse causality, omitted variables, or endogenous policy preferences.

One can repeat this intuition by contrasting alternative summaries of the HI results:

1. Interest rate fluctuations predict economic growth. In years when interest rate movements predicted slower economic growth, conflicts are more common, especially in countries with characteristics associated with higher conflict risk.
2. A period of increased conflicts, concentrated in countries with higher rates of ethno-fractionalization, occurred during a span of years when interest rates across countries were relatively high and economic growth was sluggish.

Like in the NQ case, whether one believes that the association described in the first statement establishes a causal link depends on how unlikely it would be that interest rates and conflict would be correlated in time. If all the years with high interest rates occur near each other and all the years with lots of conflict occur near each other, a spurious association between the two is not unlikely. Put differently, in panel data analysis, one must look carefully at the time series, at the sequencing of observations, and not simply assume they are independent observations over time.

As we show, the instruments, dependent variables, and endogenous regressors in both NQ and HI – and we suspect many studies in this genre - exhibit serial correlation that makes this concern salient. The NQ instrument in the baseline panel specification has a correlation with its lag >0.9, while that of the base country interest rates used by HI >0.8, and that of the conflict outcome from the NQ dataset is >.75. Our goal is to understand

how autocorrelation leads to co-trending variables, causing mistaken inference and confound identification both in terms of size and sign of expected effects, as well as how best to deal with this challenge.

2. The Underappreciated Spurious Regressions Problem in Panel IV Estimation

In order to motivate the more detailed analyses and empirical demonstrations that follow, this section provides a more general explanation of how time trends can confound panel IV estimation. We specify a general model reflective of the panel IV literature on the causes of conflict, then remind readers of the long-known mistaken inference problem caused by spurious regressions, before moving on to demonstrate that this problem can interfere with identification of true causal effects.

2.1 Spurious Regressions and Mistaken Inference in Panel IV Estimation

Consider the following highly stylized deterministic model of conflict in country-level panel data, where *i* indexes countries and *t* years[5]

$$conflict_{it} = \beta X_{it} + \psi_i \tau_t \quad (1)$$

$$X_{it} = \alpha\, conflict_{it} + \chi Z_t \quad (2)$$

The parameter of interest is the causal effect of X_{it} on $conflict_{it}$, β. This simple model captures several key points challenges to estimating the causal determinants of conflict, and the conditions under which IV estimation can help. First, the possibility that $conflict_{it}$ and X_{it} are simultaneously determined, i.e., that $\alpha \neq 0$, motivates the search for an appropriate instrument. Second, a factor exogenous to both $conflict_{it}$ and X_{it}, with

[5] In this model, an omitted time factor is responsible for correlation of the instrument Z and the error term of conflict, and reverse causality is responsible for correlation of finite sample bias across the reduced form and first stage equations. This is a plausible concern for both papers we consider and a common worry in the causes of conflict literature. Any model in which the covariance of the error term and the instrument is not constant over time will face the spurious correlation problem in the reduced form unless the non-stationarity is corrected. The endogeneity between X and conflict is what causes the finite sample bias to be signable, and could come from other sources besides reverse causation, for example in the food aid case if food aid is targeted on a variable omitted from the first stage that is also correlated with conflict. If there is no true endogeneity of conflict and X, but the reduced form has a spurious regression problem, the first stage may or may not also have a spurious regression issue, but the 2SLS coefficient would be unbiased. In that case, however, the effect of X on conflict, could simply be estimated by OLS.

both country-specific (ψ_i) and year-specific (τ_t) components may influence conflict, generating noise that might confound identification of β.[6] For example, some countries may be more prone to conflict than others, or some years may be particularly violent worldwide. Finally, an instrumental variable Z_t[7] may have a causal influence on X_{it}, but not on $conflict_{it}$, i.e., it satisfies the standard relevance and exclusion criteria for an IV. Initially, we consider an instrument that varies only in the time dimension, *t*, but we generalize this later.

Following standard IV practice, imagine that we estimate the two following regression equations by OLS:

$$conflict_{it} = \gamma Z_t + \mathrm{e_{it}} \tag{3}$$

$$X_{it} = \pi Z_t + u_{it} \tag{4}$$

Substituting equation (1) into (2) and then solving for the OLS estimates of γ and π from equations 3 and 4, we have[8]:

$$\gamma = \beta\left[\left(\frac{\alpha}{1-\alpha\beta}\right)\bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)} + \frac{\chi}{1-\alpha\beta}\right] + \bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)} \tag{5}$$

$$\pi = \left(\frac{\alpha}{1-\alpha\beta}\right)\bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)} + \frac{\chi}{1-\alpha\beta} \tag{6}$$

where $\bar{\psi}$ is the mean of ψ in the sample. Equation (5) is the reduced form parameter estimate and equation (6) is the first stage parameter estimate. The indirect least squares instrumental variable (ILS-IV) estimator (with the single instrument and without controls) is simply the ratio of these two or:

[6] Year and country fixed effects could also in principle enter as separate, uninteracted terms. Since the role of such terms can be understood by setting either effect to a constant in this model, we include only the interacted term.

[7] We focus on the case of one instrument, z_t, to highlight the role of a time series process in creating finite sample bias and because NQ use a single instrument. In practice, if multiple instruments are available, but both are subject to serial correlation or mis-specified trends, the finite sample bias in the reduced form will be a weighted average of the finite sample bias arising from both instruments. How much the additional instrument influences finite sample bias depends on the time series correlation between the instruments. For example, HI use base rate interest rates across multiple countries so that the instrument varies by country, but base country interest rates are highly correlated with the global average due to a no-arbitrage condition. We show in Appendix A that the HI results are very similar whether one uses multiple instruments or a global average interest rate.

[8] Note that we assume that Z_t is not a constant, so that $\widehat{var}(z_t) \neq 0 \; and \; \frac{\widehat{var}(Z_t)}{\widehat{var}(Z_t)} = 1$.

$$\beta_{IV} = \frac{\gamma}{\pi} = \frac{\beta\left[\left(\frac{\alpha}{1-\alpha\beta}\right)\bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)}+\frac{\chi}{1-\alpha\beta}\right]+\bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)}}{\left(\frac{\alpha}{1-\alpha\beta}\right)\bar{\psi}\frac{cov(\tau_t,Z_t)}{var(z_t)}+\frac{\chi}{1-\alpha\beta}} \quad (7)$$

At a population level, β_{IV} identifies the causal effect of X_{it} on $c_{it}$when $\frac{cov(\tau_t,Z_t)}{var(z_t)}$= 0 and $\chi \neq 0$, in which case $\beta_{IV} = \beta$.

However, in any empirical application, our sample of countries and years is inherently finite, so we estimate $\hat{\beta}_{IV}$ using an observed period T and set of countries N for which we will have a finite $\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(z_t)}$ and $\hat{\bar{\psi}}$ and consequently:

$$\hat{\gamma} = \beta\left[\left(\frac{\alpha}{1-\alpha\beta}\right)\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(z_t)}+\frac{\chi}{1-\alpha\beta}\right]+\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(Z_t)} \quad (8)$$

$$\hat{\pi} = \left(\frac{\alpha}{1-\alpha\beta}\right)\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(Z_t)}+\frac{\chi}{1-\alpha\beta} \quad (9)$$

$$\hat{\beta}_{IV} = \frac{\hat{\gamma}}{\hat{\pi}} = \frac{\beta\left[\left(\frac{\alpha}{1-\alpha\beta}\right)\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(z_t)}+\frac{\chi}{1-\alpha\beta}\right]+\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(z_t)}}{\left(\frac{\alpha}{1-\alpha\beta}\right)\hat{\bar{\psi}}\frac{\widehat{cov}(\tau_t,Z_t)}{\widehat{var}(Z_t)}+\frac{\chi}{1-\alpha\beta}} \quad (10)$$

The ILS-IV estimator will yield an unbiased and consistent estimate of β:

$$\mathrm{E}[\beta - \hat{\beta}_{IV}] = 0 \quad (11)$$

$$\operatorname*{plim}_{t\to\infty}\hat{\beta}_{IV} = \beta \quad (12)$$

$$\chi \neq 0 \quad (13)$$

Conditions (11) and (12) relate to the standard exclusion restriction, requiring that the instrument only relates to the outcome variable through its influence on X. Condition (13) is usually called the relevance condition; it requires that the instrument predicts variation in the X variable. A sufficient condition[9] for (12) in this model would be $\operatorname*{plim}_{t\to\infty}\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)} = 0$, which would imply:

$$\operatorname*{plim}_{t\to\infty}\frac{\hat{\gamma}}{\hat{\pi}} = \frac{\beta\chi(1-\alpha\beta)}{(1-\alpha\beta)\chi} = \beta \quad (14)$$

[9] An alternative condition could replace condition (8), that $\operatorname*{plim}_{N\to\infty}\hat{\bar{\psi}} = 0$. This would require that year-specific trends are on symmetric across countries, so that they are on average sufficiently close to zero for a sufficiently large sample of countries. This parallel trend assumption can be a powerful source of identification, but is not relevant to the cases we study, where exposure to conflict trends is likely asymmetric across countries.

Given these conditions, the ILS-IV estimator, when applied to a sufficiently large time series, yields a consistent estimate of the causal effect of X on conflict.

Spurious correlations can arise, however, from serial processes. For example, if τ_t and z_t each follow a pure random walk, then the second moments of the correlation coefficient $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$ are volatile, dispersed away from zero with large positive and large negative values likely. In addition, for a random walk, $\underset{t\to\infty}{\text{plim}} \frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} \neq 0$ for any T. This fact was demonstrated in simulations as early as Yule (1926), although proven analytically only recently by Ernst et al. (2017). This problem permeates panel IV estimation, complicated by the prospect of correlation between the two time series. When time series are not a pure random walk, higher persistence causes slower convergence as we add years to the time series. Standard inference then dramatically understates the likelihood of large coefficients arising by chance, leading to inflated rates of rejection of the null hypothesis that two time series variables are uncorrelated.

Taking each stage separately, one can address this problem of mistaken inference by carefully studying the time series properties of the instrument and/or of the residuals from the reduced form and first stage regressions and adjust the calculation of standard errors appropriation (e.g., using Newey-West standard errors for trend stationary variables) or transform the instrument (e.g., via first differencing, as we discuss below). The issues raised by spurious correlations of two time series are well known in the time series literature (Enders, 2008), but are not often addressed in many panel IV papers, including the ones we study here.

2.2 Consequences of Spurious Correlations for IV Estimation

Note that because the instrument only varies along the time dimension, conditions (11) and (12) turn on the correlation of two time series variables, the instrument Z_t and the time series component of the unobservable error, τ_t. This has important implications for inference in that literature and familiar designs do not routinely solve the problem.

The most well understood and commonly addressed issue is that both τ_t and z_t are processes of deterministic trends. If one omits or mis-specifies controls for these trends (for example, including a linear trend when the true trend is quadratic), then $\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)} \neq 0$, and the ILS-IV estimate of β is may be neither consistent nor unbiased. The fact that omitted trends can cause bias is well known, even if the role of misspecification is often neglected, as evidenced by the fact that papers commonly report only one trend specification – typically linear or period fixed effects common to all cross-section units– rather than a systematic approach to optimal trend specification.

Misspecified deterministic trends are not the only threat to causal identification, however. The spurious regression problem arises because of a tendency shown by Slutzky (1937) for variables that are comprised of a sum of random causes to appear to follow periodic cycles. But when two truly independent variables are both following a cycle over time, they will appear to be strongly positively correlated in periods when they both follow the upward trending part of their cycle, and strongly negatively correlated in periods when their cycles run counter to each other. Although the role of spurious regressions in IV estimation was reported by Phillips and Hansen (1990), the lessons for inference and bias seem not to have been internalized by applied researchers working with panel data. The volatility of correlations between common time series variables means that $\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)} \neq 0$ in finite samples, and will often be much larger than would be expected if both variables were truly iid. That is, even if in an infinite time series condition (12) holds and the panel IV estimator is consistent, in finite sample it will suffer bias.

Time series correlations create a problem because the possibility of simultaneity means that the time series correlation term $\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)}$ appears in both the first stage and the reduced form coefficient estimates (equations 8 and 9). To see how this can cause IV estimation to go awry, consider the case where X_{it} truly has no effect on conflict, so that the true $\beta = 0$. If controls for trends are misspecified or if the instrument and the unobservable time trend are simply spuriously correlated within sample, $\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)}$ may be large. It may, in the case of a random walk, not even converge to zero even as the time

series dimension of the panel becomes arbitrarily large. In this simple model, the reduced form, first stage, and ILS-IV estimates will be:

$$\hat{\gamma} = \bar{\psi}\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} \tag{15}$$

$$\hat{\pi} = \alpha\bar{\psi}\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} + \chi \tag{16}$$

$$\hat{\beta}_{IV} = \frac{\hat{\gamma}}{\hat{\pi}} = \frac{\bar{\psi}\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}}{\alpha\bar{\psi}\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} + \chi} \neq 0 \tag{17}$$

Although the exclusion restriction may seem plausible and the instrument is relevant ($\chi \neq 0$), the ILS-IV estimator will differ from the true $\beta = 0$. The specific influence will depend on the share of variance in the instrument explained by the true first stage χ, the spurious correlation $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$, and the source of the endogeneity α. For example, when χ is small, a high realization of $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$ will both make the irrelevant instrument strong and statistically significant, and cause us to estimate a non-zero $\hat{\beta}_{IV}$.[10] Alternatively, when the instrument is strong and endogeneity is weak, $\chi > 0$ and α close to 0, the instrument will appear strong and will have the correct first stage sign, but persistence will cause IV coefficients to be far from the true value and could have either sign depending on whether cycles are co-trending or counter-cyclical. Finally if α and χ both do not have sufficient size to dominate the first stage, the sign of the first stage could flip based on realizations of $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$, but still appear significant and strong if realizations trends are sufficiently co-cyclical, with the realizations that do reverse sign of the first stage all resulting in a $\hat{\beta}_{IV}$ that has the same sign as α. In Appendix B, we simulate this model under these three scenarios to show how persistent variables increase the risk of large realizations of $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$, with implications for first stage, reduced form, and first stage.

[10] When χ, is exactly 0, $\hat{\beta}_{IV} = 1/\alpha$, so the IV coefficient has the same sign as the endogeneity in the model.

2.3 Spurious Correlation Can Mask Weak or Irrelevant Instruments

Importantly, the finite sample bias problem exists even with a weak instrument and spurious correlation of the two time series variables can mask violation of condition (13). As is readily apparent from equation (16), $\hat{\beta}_{IV} \neq 0$ even if $\chi=0$. In that special case, when the exogenous time series instrument is in fact irrelevant, the simultaneous determination of $conflict_{it}$ and X_{it} – which motivates the IV estimation in the first place – implies that $\widehat{cov}(\tau_t, z_t)$ enters through the first stage, generating spurious relevance despite a truly irrelevant instrument, yielding $\hat{\beta}_{IV} = \frac{1}{\alpha}$, i.e., the IV estimator identifies not the true causal effect of the endogenous explanatory variable but instead the inverse of the simultaneity coefficient from equation (2). This is perhaps the most overlooked pitfall of using time series variables as instruments. We cannot trust conventional tests of the relevance condition without also checking and correcting all key variables in the first stage and reduced form estimates for time series issues such as trends and non-stationarity. Spurious correlation in the reduced form can render statistical tests of the first stage uninformative. When spurious time series correlation makes an irrelevant instrument appear relevant, the resulting distribution of IV estimates will not be centered around zero. Indeed, the ILS-IV estimate is biased in the same direction as the very source of bias that the IV was intended to solve.

2.4 Allowing for Differential Trends Using Interaction Specifications Or Related Designs

So far we have focused on simple specifications without appropriate controls for (potentially nonlinear, cyclical) trends. If variables are stationary around a trend, then correctly controlling for a trend will avoid the spurious correlation problem. Checking autocorrelation of outcomes, endogenous variables, and instruments can help diagnose a problem, but a challenge remains that unobservable trends are not defined and formal tests for stationarity such as augmented Dickey-Fuller tests are not well powered.

Given the challenge of selecting the correct trend, another approach is to interact the time series instrument with an observed characteristic w_i that varies across countries within years:

$$conflict_{it} = \beta X_{it} + \psi_i \tau_t \tag{18}$$

$$X_{it} = \alpha conflict_{it} + \chi_{int} w_i Z_t \tag{19}$$

Equation (19) is a special case of a shift-share or Bartik instrument, interacting w_i with Z_t.[11] Because $w_i Z_t$ varies in the cross section, the first stage and reduced form can be estimated with both country and year fixed effects, or equivalently by demeaning all variables before estimation. Estimating the first stage and reduced form coefficients on demeaned variables in a sample of years T and countries N gives[12]:

$$\widehat{\gamma_{fe}} = \left(\frac{1}{1-\alpha\beta}\right) \bar{\psi} \frac{\widehat{cov}\left((\psi_i - \bar{\psi})(\tau_t - \bar{\tau}), (w_i - \bar{w})(z_t - \bar{Z}) \right)}{\widehat{var}\left((w_i - \bar{w})(z_t - \bar{Z})\right)} + \frac{\beta\chi}{1-\alpha\beta} \tag{20}$$

$$\widehat{\pi_{fe}} = \left(\frac{\alpha}{1-\alpha\beta}\right) \bar{\psi} \frac{\widehat{cov}\left((\psi_i - \bar{\psi})(\tau_t - \bar{\tau}), (w_i - \bar{w})(z_t - \bar{Z}) \right)}{\widehat{var}\left((w_i - \bar{w})(z_t - \bar{Z})\right)} + \frac{\chi}{1-\alpha\beta} \tag{21}$$

$$\hat{\beta}_{feIV} = \frac{\widehat{\gamma_{fe}}}{\widehat{\pi_{fe}}} = \frac{\frac{\widehat{cov}((\psi_i - \bar{\psi}), (w_i - \bar{w}))\widehat{cov}((\tau_t - \bar{\tau}), (z_t - \bar{Z}))}{\widehat{var}\left((w_i - \bar{w})(z_t - \bar{Z})\right)} + \beta\chi}{\alpha \frac{\widehat{cov}((\psi_i - \bar{\psi}), (w_i - \bar{w}))\widehat{cov}((\tau_t - \bar{\tau}), (z_t - \bar{Z}))}{\widehat{var}\left((w_i - \bar{w})(z_t - \bar{Z})\right)} + \chi} \tag{22}$$

The potential advantage of this specification is that it introduces scope for the IV to be identified, consistent, and unbiased through an assumption related to the cross sectional variables rather than relying on the time series dimension alone, since if $\widehat{cov}((\psi_i - \bar{\psi}), (w_i - \bar{w}))= 0$, $\beta= \hat{\beta}_{feIV}$. In this model, the assumption necessary to identify the causal effect is clearer, however. If unobserved serial shocks affect conflict differently across countries, relative exposure to this variable must be uncorrelated with the variable interacted with the time series instrument. This assumption would be violated if, for

[11] We emphasize that these are merely a special case of shift-share instruments. The more general shift-share instrument is a weighted sum of multiple exogenous shocks, where the weights vary among cross-sectional observations. The instruments in our case – and in HI, NQ and most of this literature – use a single shock that varies in the time series domain exclusively, to which cross-section units are differentially exposed. This is akin to an instrument continuous difference-in-differences regression. With a single shock, the shift-share IV cannot leverage exogenous variation across multiple shocks, which gives the method much of its power (see Borusyak et al. 2018 and Adao et al. 2019 for more details). Specifically, the basic identifying assumption of Borusyak et al. (2018) - that Z_t is as-good-as-random over time - no longer holds whenever the regression includes time fixed effects. Moreover, the standard errors of Adao et al. (2019), which correct for units' common exposure to shocks, no longer work in the case of a single shock. We thank an anonymous referee for flagging these important distinctions.

[12] In these derivations we also make the mild assumptions that $\alpha\beta \neq 1$ and $\widehat{var}((w_i - \bar{w})(z_t - \bar{Z})) \neq 0$.

example, the countries most targeted for food assistance are more exposed to global conflict trends, as seems quite plausible.

3. Panel IV Methods Without Interactions in Empirical Examples

We now illustrate the issues with panel IV with reference to two prominent papers that study the causes of conflict: NQ and HI. These applications demonstrate how persistence influences the proposed IV strategies and how spurious regressions generate mistaken conclusions in real data. We supplement these below with fully controlled simulations that explicate more precisely the underlying mechanisms behind the identification problem in panel IV estimation.

We begin by first ignoring the interacted instrument construction behind the HI and NQ papers to focus attention on the time series properties of the variables of interest, to show how these properties generate the spurious regressions problem on which we focus. We turn to the IV with interactions in the next section. In the simplest form, both HI and NQ estimate a central relationship of the following type:

$$Conflict_{it} = \beta X_{it} + \epsilon_{it} \tag{23}$$

In NQ, X_{it} is the quantity of US wheat food aid shipped to country i in year t; in HI, X_{it} is the growth of real GDP in country i from year t-1 to year t. In both cases, X_{it} is likely endogenous to conflict, even if one controls for country and year fixed effects or other observable control variables. In the food aid case, US government policy explicitly states that food aid should be sent to countries experiencing active conflict or perceived to be at risk of conflict.[13] Such a policy likely creates upward bias when estimating β by OLS because any factors that increase the risk of conflict that are observed by the US government but not controlled for in the regression would be positively correlated with both X_{it} and conflict. Another hypothesis is that, despite stated policy, less food aid gets delivered to countries at higher risk of conflict because of logistical difficulties or the

[13] "Food for Peace saves lives, reduces suffering and *supports the early recovery of people affected by conflict* and natural disaster emergencies through food assistance" (https://www.usaid.gov/who-we-are/organization/bureaus/bureau-democracy-conflict-and-humanitarian-assistance/office-food, emphasis added).

higher costs of working in conflict locations. In HI and many other papers in the literature on conflict and development (reviewed by Ray and Esteban 2017), β is potentially biased downwards by reverse causality if active conflict dampens economic activity.

So both HI and NQ naturally turn to IV estimation, proposing a variable, Z_t, that is correlated with X_{it} and uncorrelated with conflict except through X_{it}. In NQ, Z_t is lagged (i.e., year t-1) total wheat production in the US. In HI, Z_t is the short-term nominal interest rate of the base country to which country *i*'s exchange rate is most closely tied. The concern is that a no-arbitrage condition implies cointegration of interest rate movements across countries, meaning that interest rate movements are mostly explained by average movements. [14] To highlight this problem and show where spurious regression enters into panel IV, we substitute the HI instrument with the global average real interest rate so that instrument Z_t varies only in the time series dimension, not in the cross-section of countries and show that the results are the same as reported in their paper.[15]

In simplified form, both papers estimate the effect of their endogenous variable on conflict through a two stage least squares (2SLS) procedure consisting of the two regressions:

$$Conflict_{it} = \gamma^{base} Z_t + \theta_i + \rho_{ir} t + \mu_{it} \quad (24)$$

$$X_{it} = \pi^{base} Z_t + \Theta_i + \mathrm{P}_{ir} t + \eta_{it} \quad (25)$$

Equation 25 is the first stage, estimating the causal effect of the exogenous instrument, Z_t, on the endogenous regressor, X_{it}. Equation 24 estimates the reduced form relationship between conflict and the instrument, Z_t. These equations can include controls for countries and a time trend interacted with a dummy variable for the world region r of which country *i* is a member, but since the instrument only varies annually in the time series, they cannot include year fixed effects. The indirect least squares (ILS) IV estimate, the ratio of the reduced form estimate over the first stage coefficient estimate, $\widehat{\gamma^{base}}/\widehat{\pi^{base}}$, represents the 2SLS estimate.

[14] HI follow Shambaugh (2004) in classifying countries whose currencies are not explicitly pegged to another country's currency via a fixed exchange rate.
[15] In Appendix A we demonstrate that this simplification has no qualitative effect on our results relative to using the vector of base interest rates HI use.

3.1 Ignoring the Time Series Nature of the Data

We begin by replicating the NQ (HI) analyses, using the same panel data including 125 (97) non-OECD countries over 36 (34) years, with the binary dependent variable of conflict status, which equals one if a country experienced more than 25 battle deaths in a year, the endogenous regressors of quantity of wheat food aid delivered to country *i* by the US (year-on-year GDP growth in *i*), the instruments – lagged US wheat production (global real interest rates) – and a rich set of characteristics of countries and years that the original authors use as controls.[16] When one looks at simple scatter plots of data, ignoring the temporal sequencing of observations, the panel IV identification strategy seems to work. Figure 1a shows the correlation between real interest rates and conflict, the reduced form relationship in HI. Interest rates and conflict covary positively. Figure 1b shows the negative first stage relationship with the endogenous variable. Since the IV estimate is just the reduced form divided by the first stage, we know that the ILS/2SLS estimate of GDP growth on conflict, instrumenting for growth with interest rates, will be negative, i.e., that GDP growth is associated with less conflict.

[16] The main variables of interest for NQ are taken from the UCDP/PRIO Armed Conflict Dataset Version 4-2010 (conflict), the Food and Agriculture Organization's (FAO) FAOSTAT database (food aid deliveries), and the USDA (wheat production). In replicating both papers, we accessed the NQ replication file included with the publication in the *American Economic Review* (available online at https://www.aeaweb.org/articles?id=10.1257/aer.104.6.1630) to ensure that we used the identical version of these data as NQ. These data are described in further detail in the original NQ paper. Because the HI paper does not include a publicly available replication file, the real interest rate variable is taken from the World Development Indicators (World Bank 2018) and merged into the NQ dataset. We are therefore explicitly not attempting to replicate HI's numeric estimates, just their procedure using similar data.

Figure 1a: Conflict and real interest rates

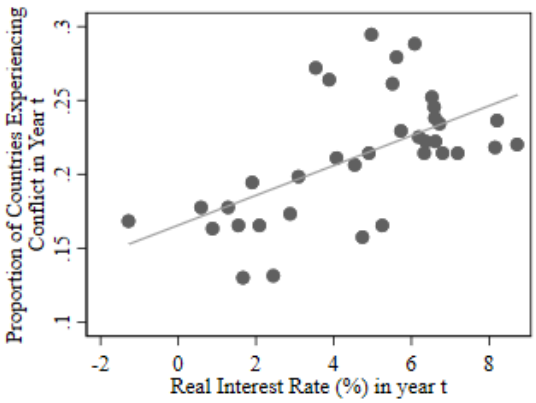
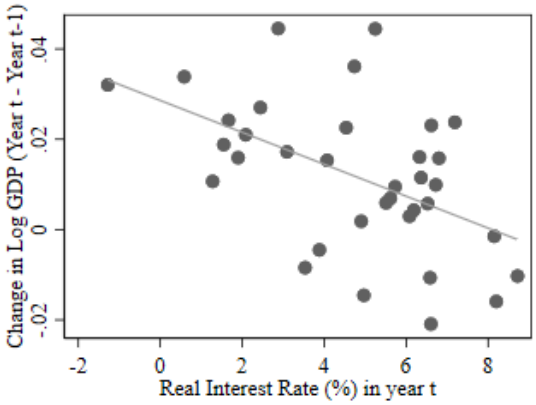


Figure 1b: GDP growth and real int. rates



Notes: Conflict and GDP are from NQ dataset as posted by the American Economic Review, including 127 countries in 1971-2006. Real interest rates from World Development Indicators, (World Bank, 2018). All figures are the raw data, unadjusted for controls.

Figure 2a:

Conflict and lagged US wheat production

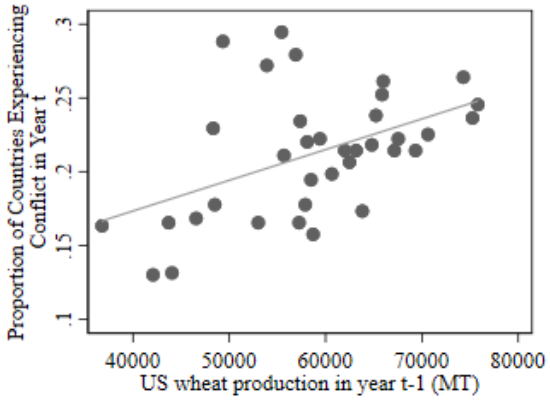
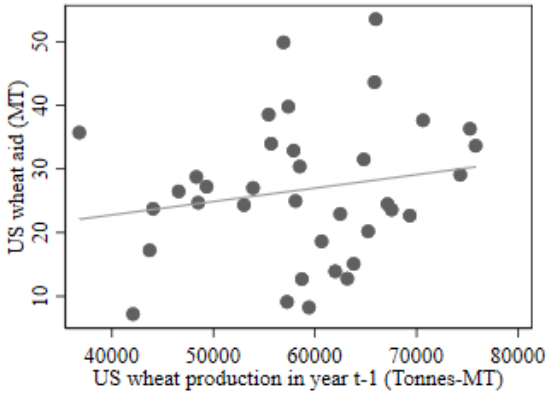


Figure 2b:

Food aid and lagged US wheat production



Notes: Data are from NQ dataset as posted by the American Economic Review, including 127 countries in 1971-2006. All figures are raw data, unadjusted for controls.

Similarly, Figure 2a shows the positive reduced form relationship in NQ, between conflict and lagged US wheat production, while Figure 2b shows the positive first stage relationship between lagged US wheat production and wheat food aid shipments. Since both the first stage and reduced form relationships are positive, the ILS/2SLS estimate of US food aid, instrumented by lagged wheat production, on recipient country conflict is

necessarily positive as well, suggesting that food aid is positively associated with (prolonged) conflict.

3.2 Assessing Trends in the Data

The problem with the estimation strategy above is that the sequencing of observations plays no role in the analysis, although the data come from specific time series. One could scramble the time series observations without changing the plots in Figures 1 and 2 and the parameter estimates based on the relationships depicted in them at all.

Figure 3 displays the actual trends in the time series, shown in means and estimated nonparametrically by lowess, in conflict (upper left panel), US wheat production (upper right panel), interest rates (lower left panel) and a fourth variable, global audio cassette tape sales (lower right panel). We chose the audio cassettes variable specifically because it is obviously spurious but exhibits a clear, nonlinear trend. [17] No coherent, credible mechanism exists that causally links audio cassette tape sales to conflict, real interest rates, or US food aid shipments.[18] Conflict, US wheat production, and global real interest rates all followed the same inverted-U trend over the sample period as do global audio cassette tape sales.

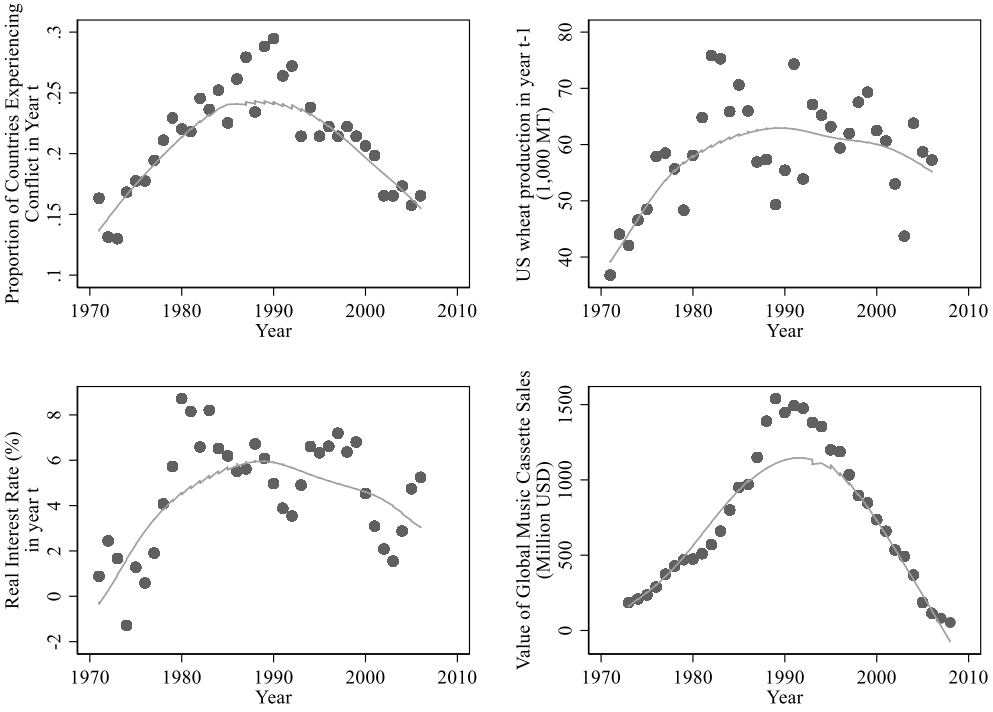
The simple reduced form estimates in Table 1 confirm what one can immediately infer from visual inspection of the plots of the time series: strongly positive and statistically significant correlations between the dependent variable of interest, conflict, and each of the other three candidate instrumental variables. It does not matter whether the instrumental variable is plausible, like real interest rates or lagged US wheat production, or obviously spurious, like global audio cassette sales. The reduced form is strong and positive regardless. This underscores an important point widely underappreciated in panel IV

[17] The global audio cassette sales data come from IFPI (2009). If one tries enough variables, one can always find a spurious variable that is correlated with the others. We chose this variable because it shows the role of trends in creating an apparently significant association in both the first stage and reduced form.

[18] Finding a spurious correlation is not sufficient to show that a given IV strategy is invalid. In finite series, given enough variables one could always find through multiple hypothesis testing an obviously unrelated variable that returns a spuriously non-zero correlation. We chose this one because simple visual inspection of the data immediately reveals the source of the spurious association with conflict.

estimation. If the outcome of interest exhibits a strong trend, then any variable that exhibits a similar (opposing) trend will generate a statistically significant, positive (negative) reduced form relationship, whether or not the instrument is spurious or truly causal. How can we rule out the possibility that plausible instruments like lagged US wheat production or global real interest rates are not spuriously correlated with the outcome of interest just like the clearly spurious instrument, global audio cassette tape sales?

Figure 3: Underlying trends in the conflict and instrumental variables



Notes: Conflict and wheat data from NQ (from American Economic Review repository), include 127 countries, 1971-2006. Real interest rates from World Development Indicators, (World Bank, 2018) and cassette tape sales from IFPI (2015). All figures are unadjusted for controls. Trends are estimated by lowess. Dots show yearly averages.

The relationship we care about is not the reduced form, but rather the relationship between the outcome (conflict) and the potentially endogenous explanatory variable (shipments of food aid or GDP growth). A reduced form relationship between an outcome and an instrument is only one criterion to check in determining the validity of the IV strategy. The other is the first stage correlation to validate the relevance of the instrument.

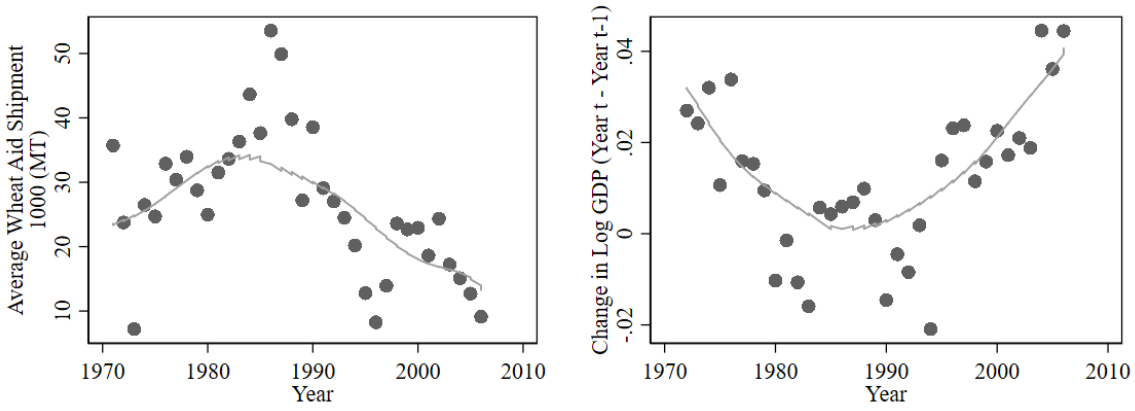
We know from Figures 1 and 2 that real interest rates are associated with GDP growth and that lagged US wheat production is correlated with food aid shipments. But in those figures, time played no role. Figure 4 displays the trends in the endogenous regressors of interest: wheat food aid in panel 4a and GDP growth in panel 4b. Both variables also show a strong trend, inverted-U in the case of wheat food aid shipments, just like the outcome variable and candidate instruments displayed in Figure 3, and U-shaped in the case of real GDP growth, counter-cyclical to the plots previously displayed.

Table 1: Reduced form estimates between conflict and candidate instruments

	Dependent variable = incidence of war (of any type)		
VARIABLES	(1)	(2)	(3)
Global real interest rate	0.01082		
	(0.00345)		
Lagged US wheat production		0.00245	
		(0.00076)	
Global music cassette sales			0.08196
			(0.02162)
Observations	4,161	4,161	3,964
R^2	0.482	0.481	0.494

Note: All regressions include country fixed effects and year trends interacted with one of six geographic regions defined by the World Bank as estimated by NQ. Robust standard errors are clustered at the country level as in NQ and HI. Conflict, and US wheat production are taken from the NQ dataset, interest rates from the World Development Indicators, and music cassette sales from IFPI (2015).

Figure 4a: US wheat food aid shipments *Figure 4b: GDP growth rates*



Notes: Data from NQ (from American Economic Review repository), include 127 countries, 1971-2006. All figures are raw data, unadjusted for controls. Trends are estimated by lowess. Dots are yearly averages.

Given that our candidate instruments all have inverted-U trends, Figures 3 and 4 tell us what we already knew from Figures 1 and 2, that interest rates will be negatively correlated with GDP growth and that lagged US wheat production will be positively correlated with food aid shipments in a given year. It is less obvious, however, at least until one compares multiple variables' trends, that any of several candidate instruments and variables with common or mirror-image trends can generate significant panel IV estimates of the relationship of interest, whether or not the instruments are spurious.[19] A common trend among the dependent, endogenous explanatory, and instrumental variables means that spurious and truly causal relationships will exhibit identical patterns, calling into question the causal identification. As shown in section 2, the spurious correlation in the time series dominates even the true irrelevance of a candidate instrument, generating biased panel IV estimates.

Table 2 reinforces this concern, demonstrating that co-trending instruments serve as strong substitutes for one another. Instrumenting for GDP growth or US food aid shipments with any of the three candidate instruments – global real interest rates, lagged US wheat production, or global audio cassette sales – yields remarkably similar coefficient

[19] We use HI and NQ precisely to illustrate this in the case of both common and opposite cycles.

estimates that are always highly statistically significant. Indeed, for the food aid regressor NQ study, the most precise 2SLS estimate comes from using the audio cassette tape sales instrument that is most obviously spurious. The multiple candidate instruments raise a concern that some omitted cyclical variable – the rise and fall of Reagan-Thatcher policies? El Nino Southern Oscillation climate cycles? – may account for the observed correlations.[20]

Table 2: Co-trending instruments as substitutes for one another

	Dependent variable = incidence of war (of any type)					
	(1): R	**(2): W**	**(3): C**	**(4): R**	**(5): W**	**(6): C**
GDP growth	-2.97560	-3.12900	-3.49071			
	(1.07478)	(1.27973)	(1.10815)			
US food aid (tons)				0.00844	0.00506	0.00848
				(0.00834)	(0.00332)	(0.00309)
Observations	4,015	4,015	3,917	4,161	4,161	3,964

Note: Column headers indicate the instrument used. R= real interest rates, W = lagged US wheat production, C = cassette tape sales. All regressions include country fixed effects and year trends interacted with one of six geographic regions defined by the World Bank as estimated by NQ. Robust standard errors are clustered at the country level as in NQ and HI. Conflict, and wheat production are taken from the NQ dataset, interest rates from the World Development Indicators, and music scales from IFPI (2015).

First stage inference tests do not help us identify the spurious correlation. Using cassette sales as an instrument for previous year's US wheat production or real interest rates, the first stage t-statistics are 85 and 109, respectively, and the Kleibergen-Paap weak instrument F-statistics of 32.4 and 10.1, exceed the standard threshold value of 10.

In finite samples, one can always find a spurious variable that is highly correlated with the outcome variable. The fact that global audio cassette sales are correlated with conflict does not mean that more food aid or slower GDP growth do not cause conflict. Rather, it hints at the challenges to making valid inference that arise from spurious correlation in time series variables.

[20] This table also raises a publication bias concern. One could imagine constructing an IV strategy using global real interest rates to instrument for food aid deliveries. The standard IV analysis would suggest that interest rates have a strong first stage.

3.3 How Correlated Cycles Affect Panel IV Inference: A Monte Carlo Analysis

We next demonstrate that the spurious result is not unique to a single variable, using Monte Carlo simulation to show how autocorrelation in time series variables can cause the mistaken inference, finite sample bias, and weak instruments problems we explained algebraically earlier. We draw on the time series literature dating back at least to Yule (1926), Slutzky (1937), Granger and Newbold (1974), Phillips (1986), Phillips and Hansen (1990), and Phillips (1998), all of whom found correlated errors can cause standard inference tests to suggest spurious statistical significance.

We begin by simulating an instrument that follows a random walk process.[21] Specifically, in each round we implement the following procedure on equations (1)-(4) from above, mimicking the NQ study except that we replace lagged US wheat production, their instrument, with a manufactured random variable that explicitly follows a nonstationary, random walk process. By construction, this is an irrelevant instrument. The simulation protocol is:

1. Define an instrumental variable Z_t that takes a value of 100 in year 1.
2. In each subsequent year, there is a random shock that is uniformly distributed, $q_t \sim$ U(-.5,5). In year t, $Z_t = Z_{t-1} + q_t$. Therefore, any given year Z's expected value, $E[Z_t]=100$, but the realized value, Z_t, will fall above or below its expected value based on the prior sequence of innovations in q_t. From year 1 onward, Z_t follows a random walk.[22]
3. In years 1-36[23], holding conflict, food aid flows, and all of NQ's controls from their baseline specification constant across iterations, we estimate the first stage, reduced form, and 2SLS equations from the baseline model reported by NQ,

[21] In Appendix B we show in a fully controlled simulation that the basic patterns hold for a range of serial correlation parameters and for a relevant instrument as well. This demonstrates empirically that our core results do not depend on either a weak instrument nor on difference stationarity. The simple framework we use here allows us to show that the problems we highlight need not arise from any specific omitted variable, deterministic trend, or weak instrument. The problem arises simply from the smooth dynamics of the instrument when some information from past realizations persists.

[22] In Appendix E, we show that this result is not specific to a random walk by adding a coefficient ρ to $Z_t = \rho Z_{t-1} + q_t$ and varying the size of ρ from 0 to 1. The problem of volatility increases as ρ approaches 1.

[23] We use 36 periods simply to replicate the duration of the sample used in NQ's estimation. This is inherently arbitrary, but reflects a period T that could and does appear in literature.

substituting the Z_t variable described above as the instrument for food aid rather than lagged US wheat production. Everything else stays exactly the same as in NQ; we use the data and code from their replication package.

4. Repeat steps 1-3 1,000 times, saving the coefficient estimates on Z_t, the associated p-values and KP F-statistics for weak instrument tests in the first stage, reduced form, and 2SLS equations.

The upper left panel of Figure 5a plots the distribution of the π^{sim} coefficients estimated in each of 1000 replications of the following first stage regression:

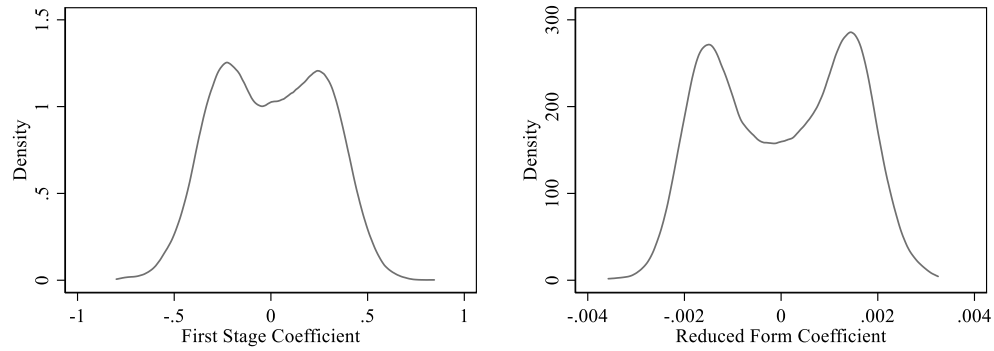
$$Aid_{it} = \pi^{sim} Z_t + \boldsymbol{Controls_{irt}} \Pi^{\text{sim}} + \theta_i^{sim} + \rho_{ir}^{sim} t + \eta_{it}^{sim} \quad (26)$$

In expectation, Z_t is uncorrelated with Aid_{it}, i.e., E(π^{sim})= 0. But the distribution exhibits a multi-modal pattern first reported by Yule (1926).[24] While the mean of π^{sim} across simulated draws of the data set indeed equals zero, the mode diverges *away* from the expectation, so that extreme values arise more often than values close to the true population parameter, zero. This illustrates the mistaken inference problem that arises due to spurious correlation of the time series.

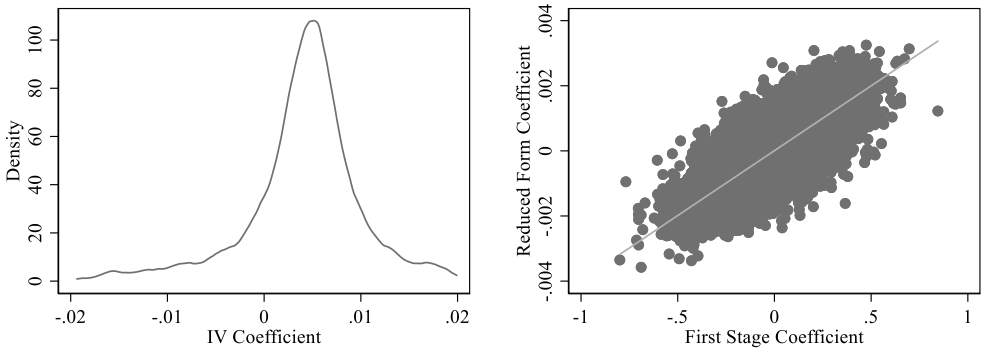
[24] Yule's (1926) empirical finding remains a subject of analytical research in statistics. Ernst et al. (2017) recently proved the result that the distribution of estimated correlation coefficients between two independent time series will be highly dispersed. There do not yet appear to be analytical results for the panel data or instrumental variables estimation cases, however. The empirical simulation methods we use appear to remain the state of the art currently.

Figure 5: Monte Carlo panel IV estimates with positively co-trending variables

5a: first stage estimate distribution *5b: reduced form estimate distribution*



5c: 2SLS coefficient estimate distribution *5d: reduced form and first stage estimates*



Notes: Distributions from simulating a fake instrument Z_t 1,000 times. Conflict outcome, food aid allocations, and controls are taken from the NQ replication dataset and the NQ baseline specification. Panel a plots the estimated $\widehat{\pi^{sim}}$ from equation (26) from each simulated dataset. Panel b is the distribution of $\widehat{gamma}^{sim}$ from equation (27), and panel c is the IV coefficient.

Figure 5b shows the distribution of coefficient estimates from the reduced form equation:

$$Conflict_{\mathrm{it}} = \gamma^{sim} Z_{it} + \boldsymbol{Controls_{irt}} \Gamma^{\mathrm{sim}} + \Theta_i^{sim} + \mathrm{P}_{ir}^{sim} t + \mu_{it}^{sim} \quad (27)$$

Not surprisingly, since we already know that the conflict variable cycles too, this distribution also exhibits Yule's "nonsense correlation" problem. The mass of estimated coefficients again occurs *away* from zero, even though the coefficient estimate converges to zero in expectation. Conventional significance tests of the reduced form will also understate the p-value of the estimated relationship.

Note that the Yule-Slutzky spurious regressions issue in either the first stage or the reduced form regressions alone is a problem of mistaken inference, not bias. The estimated $\widehat{\pi^{sim}}$'s in Figure 5a and $\widehat{\gamma^{sim}}$'s in Figure 5b center around zero, confirming that across "experiments" $E[\widehat{\pi^{sim}}] = \widehat{\gamma^{sim}} = 0$ in both cases. When focusing only on one or the other equation, the issue is that standard inference tests are based on the assumption that π^{sim} has a unimodal (typically, normal) distribution. Conventionally computed p-values will therefore understate the probability that $\widehat{\pi^{sim}}$ or $\widehat{\gamma^{sim}}$ is at least as far from the zero null value as the observed value when the actual sampling distribution is multi-modal, thereby artificially inflating the estimated statistical confidence that a relevant relationship exists. We may take small comfort then from the fact that if we repeat the study enough times, we will eventually get the right answer for the reduced form and the first stage relationships, as conventional inference tests will not reveal within a given study which results are valid.[25]

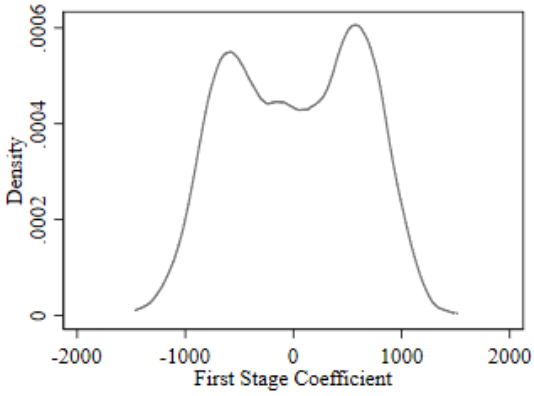
The greater concern is that the unbiasedness that holds for the OLS estimate estimated in the first stage or in the reduced form equation does not hold for the 2SLS/ILS estimate. The empirical distribution of the 2SLS estimate, shown in Figure 5c, is clearly positively biased and not centered around zero, as researchers implicitly assume would be true the case if instruments are irrelevant. The reason is evident in Figure 5d. The first stage and reduced form estimates from the same regression are positively correlated. This occurs, quite predictably, because the conflict and food aid variables follow the same inverted-U cycles. This positive correlation in trends generates positive bias in the IV estimate of interest, arising purely due to the spurious regressions problem. As we saw in the simple model analyzed algebraically earlier, when we find a spurious correlation in the reduced

[25] The concern about consistency for increasing t within a given sample is practically relevant here. In simulations on increasingly long segments of the observed data, we find that the average of coefficients does not uniformly increase or decrease with longer T samples, reported in Appendix C. As Yule (1926, pp. 12-13) put it, "[b]e it remembered, we have taken a fairly long sample [to establish the independence of two cycling variables]… if the complete period were something exceeding, say, 500 years, it is seldom that we would have such a sample at our disposal." If a cycling variable only finishes its cycle once every 500 years, we may need 500 years of data to reveal the true association with another cycling variable. To make this situation worse, if the cycling is a result of random processes, as described by Slutzky (1937), the length of time needed to "finish a cycle" may not be known, because it does not result from any model other than the structure of the unobserved error process. See Appendix C for a more detailed exploration of the issue of consistency in t.

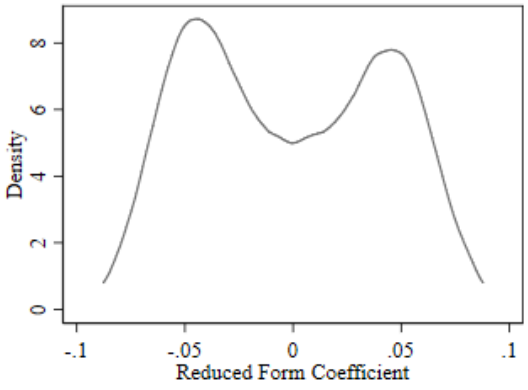
form, endogeneity between conflict and aid creates correlation in these variables. So when spurious correlation appears in one step of the IV process, the odds that it arises in the other step are high, and jointly they bias the parameter estimate of interest.

Figure 6: Monte Carlo panel IV estimates with negatively co-trending variables

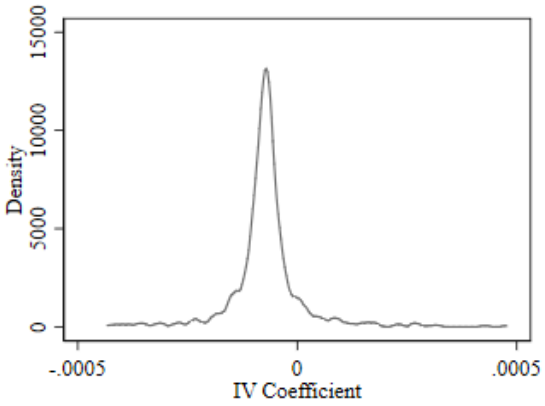
6a: first stage estimate distribution



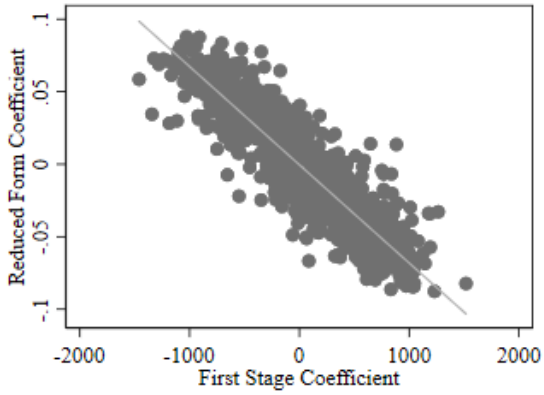
6b: reduced form estimate distribution



6c: 2SLS coefficient estimate distribution



6d: Reduced form and first stage estimates



Notes: Distributions from simulating a fake instrument Z_t 1,000 times. Conflict outcome and controls are taken from the NQ replication dataset and the NQ baseline specification; interest rates are from WB 2018. Panel a plots the estimated $\widehat{\pi^{sim}}$ from equation (26) from each simulated dataset. Panel b is the distribution of $\widehat{gamma}^{sim}$ from equation (27), and panel c is the IV estimates.

Figure 6 repeats the exercise, now using GDP growth rather than food aid as the endogenous X variable, following HI. The distribution of reduced form coefficient

estimates in Figure 6a again shows the now-expected bimodal pattern with a disproportionate incidence of coefficient estimates farther from zero than near zero. The distribution of first stage coefficient estimates from regressing GDP growth on the spuriously generated random walk variables shows this same bimodal pattern of spurious correlation in Figure 6b that we saw in Figure 5b.

The difference between the NQ (food aid) and HI (GDP growth) models is apparent in the bottom two panels of Figures 5 and 6. In Figure 5c (6c) we find that the Monte Carlo analog to the NQ (HI) estimates are positively (negatively) biased and the reduced form and first stage coefficient estimates are positively (negatively) correlated in the case where the endogenous regressor and outcome variable co-trend (counter-)cyclically.

If both $\widehat{\gamma^{sim}}$ and $\widehat{\pi^{sim}}$ are estimated by spurious correlations, and the two may be correlated, as we saw in Figures 5 and 6, can we trust the 2SLS IV estimate truly identifies the causal effect of the endogenous regressor? Clearly not. Although $E[\widehat{\pi^{sim}}] = 0$ and $E[\widehat{\gamma^{sim}}] = 0$ for the average replication of this hypothetical IV experiment, $E[\widehat{\gamma^{sim}} / \widehat{\pi^{sim}}] \neq 0$ unless $\widehat{\pi^{sim}}$ and $\widehat{\gamma^{sim}}$ are uncorrelated, which is unlikely given spurious correlation between cycles of time series variables that exhibit persistence.

Appendix B generalizes these empirical results using a fully controlled system of equations of known parameterization. In each model, the true causal parameter of interest, β from equation (1), equals zero. In simulation models 1, 2, and 3, we vary the degree of serial correlation, the relative strength of the instrument and of the endogeneity in the first stage, as reflected in the coefficients χ and α, respectively, from equation (2). Within each model, we simulate under varying degrees of persistence of the random innovations in the time series, ρ, from the set $\rho = \{0.0, 0.1, 0.5, 0.6, 0.9, 1.0\}$, i.e., ranging from iid through a fully I(1) time series. In model 1, with the case of a strong instrument and weak endogeneity.

The fully simulated results echo our previous findings. The parameter estimates for the first stage and reduced form are unbiased and their sampling distributions are reasonably behaved when $\rho=0$, but their sampling distributions become increasingly diffuse as ρ increases. This replicates the canonical spurious correlation result from the

time series literature. Model 2 makes the endogeneity prominent. The resulting first stage and reduced form equation coefficient estimates are unbiased but suffer from mistaken inference as ρ increases. But because the first stage and reduced form coefficient estimates are strongly, spuriously correlated a pronounced bias emerges in the IV coefficient estimates, in the same direction of the reverse causality the IV is meant to resolve. Because of that correlation, and unlike the sampling distributions of the first stage and reduced form estimates, the sampling distribution of the IV estimates becomes more rather than less concentrated as ρ increases, resulting in firmer erroneous rejection of the null. This is the finite sample bias problem previously unrecognized in panel IV estimators. When simultaneity between the outcome variable and the endogenous explanatory variable is prominent relative to the strength of the instrument in explaining the endogenous explanatory variable, the IV estimates become biased in the direction of the reverse causality the IV is meant to resolve, and more pronouncedly so as persistence increases.

Finally, in Model 3, we show a case where endogeneity and the strength of the first stage are sufficiently balanced that the extent to which the spuriously correlated trends overwhelm the sign of the true first stage and reduced form equations depends on the degree of persistence. In all three cases, removing the serial correlation through first differencing helps address the identification problem, reducing the volatility of the IV coefficient in model 1, reducing the volatility of the first stage in model 2, and both reducing and shifting the distribution of coefficients toward the true value of β.

3.4 Addressing the Common Cycles Problem

The fundamental issue with inference and identification in panel IV estimation is the strong assumption that $cov(\epsilon_{it}, \epsilon_{jt})$ is constant for all t, which may not be appropriate if realizations of either the outcome variable or the endogenous X variable depend on past realizations. A common strategy to address this concern is to control for past realizations in the regression equations. For example, NQ report a robustness check where they add past realizations of conflict as controls. The two equations of the 2SLS framework then become

$$Conflict_{it} = \gamma_1^{ldv} Wheat_{t-1} + \gamma_2 Conflict_{it-1} + \boldsymbol{Controls_{irt}}\Gamma^{\text{sim}} + \Theta_i^{sim} + \mathrm{P}_{ir}^{sim} t + \mu_{it} \quad (28)$$

$$Aid_{it} = \pi_1^{ldv} Wheat_{t-1} + \pi_2 Conflict_{it-1} + \boldsymbol{Controls_{irt}}\Gamma^{\text{sim}} + \Theta_i^{sim} + \mathrm{P}_{ir}^{sim} t + \eta_{it} \quad (29)$$

This specification allows for correlation between conflict in periods t and t-1. If US wheat production (*Wheat*) is exogenous and iid over years and conflict is iid over time conditional on the previous year's conflict, then this obviates the spurious regression problem in the reduced form regression of conflict on US wheat production. But the reduced form equation is only one part of the 2SLS framework. If aid flows or wheat production are also nonstationary, as appears true in Figure 4a, then the first stage regression of aid on conflict still risks the spurious regression problem.

In order to explore the effects of trying to control for prospective serial correlation in the outcome or endogenous explanatory variable, we expand the Monte Carlo simulation described above to include three additional specifications:

(i) LDV: We control for the lagged value of the dependent variable (*Conflict*) and generate the ILS/2SLS estimates, as before;

(ii) LIV: We control for the lagged value of the independent variable (*Aid*) and generate the ILS/2SLS estimates, as before;

(iii) 1st Diff: we take first differences of all variables (*Conflict, Aid*, and *Wheat*) and generate the ILS/2SLS estimates, as before. Note that because the manufactured, irrelevant instrumental variable follows an I(1) process, first differencing will necessarily generate an iid process. This will not be true more generally, when one does not know the true nature of the nonstationary process the variable follows.

The first differences specifications estimate the following reduced form and first stage equations

$$\Delta Conflict_{it} = \gamma_1^{diff} \Delta Z_t + \boldsymbol{Controls_{irt}}\Gamma^{\text{sim}} + \Theta_i^{sim} + \mathrm{P}_{ir}^{sim} t + \mu_{it}^{diff} \quad (30)$$

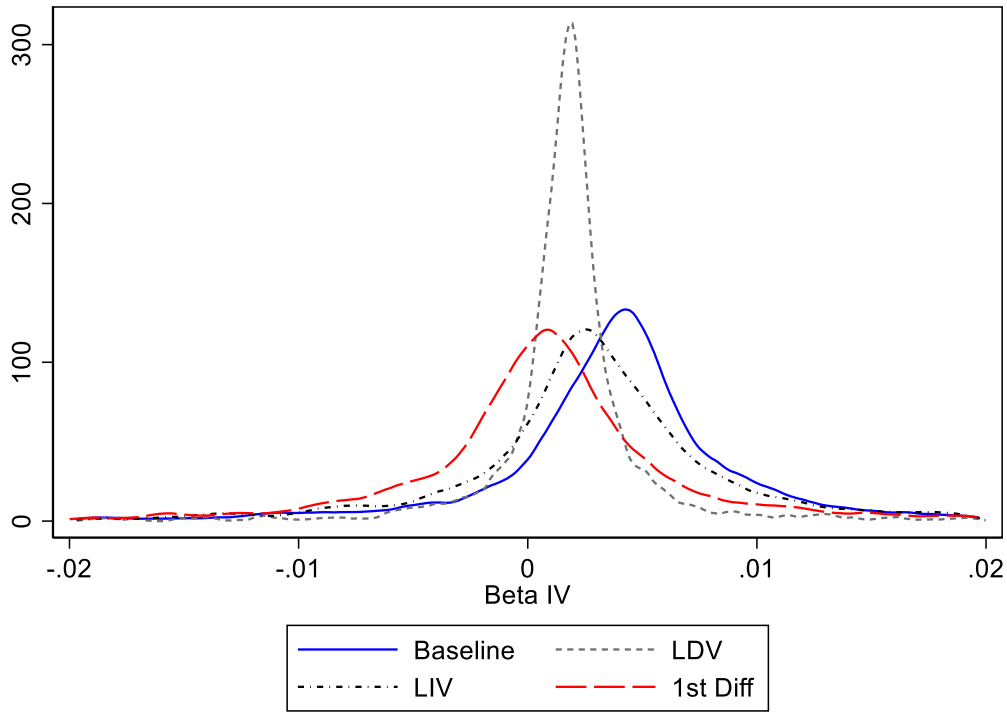
$$\Delta Aid_{it} = \pi_1^{diff} \Delta Z_t + \boldsymbol{Controls_{irt}}\Gamma^{\text{sim}} + \Theta_i^{sim} + \mathrm{P}_{ir}^{sim} t + \eta_{it}^{diff} \quad (31)$$

For each simulation, we plot the distribution of $\gamma_1^{ldv}/\pi_1^{ldv}$ parameter estimate for 1,000 draws of the simulation along with the distribution from the baseline specification as above (Figure 7). Controlling for only the LDV or the LIV does not eliminate the bias from spurious regressions. The distributions of $\gamma_1^{ldv}/\pi_1^{ldv}$ when controlling for the lagged LDV

or lagged LIV are both centered above zero. This is unsurprising since the LDV and LIV specifications only correct persistence in either the outcome or the independent variable, but the problem could be with either or both. The standard error of the distribution is smaller for the LDV case than for the baseline, meaning that depending on the relative reduction in error or mean bias, including the lagged LDV could actually increase the odds that one mistakenly reports a statistically significant non-zero relationship due to the use of a spurious instrument.

As reflected in Figure 7, only a first differences specification does not on average return an estimated positive effect of aid on conflict.[26] This is intuitive because first differencing exactly corrects for the known I(1) process of the manufactured instrument.

Figure 7: Distributions of 2SLS parameter estimates



Notes: Distributions from simulating a fake instrument Z_t 1,000 times. Outcome variables and controls are taken from the NQ replication dataset and the NQ baseline specification.

[26] In appendix E, we expand this check for instruments governed by different autocorrelation parameters. The effect is strongest for a random walk is stronger when autocorrelation is higher, but is still apparent for example when autocorrelation of the instrument is above .6, well above the NQ and HI applications we study.

Given this finding, we implement the NQ 2SLS estimation strategy to estimate the coefficient of aid on conflict – in an uninteracted model not yet accounting for shift-shares – taking first differences across years as in equations 30-31. We compute standard errors clustering at the country level as in both NQ and HI. The resulting coefficient estimates reported in Table 3 are similar in magnitude to those originally reported by NQ, but in the opposite direction – i.e., suggesting a negative effect of aid on conflict – and statistically insignificant. Correcting for prospective nonstationarity in the time series completely overturns NQ's headline result.

Table 4 replicates this exercise for the HI 2SLS estimation of the effect of GDP growth on conflict using the global real interest rate to allow us to use all countries in the NQ dataset. The coefficient estimate on GDP growth is likewise not statistically significant in any specification and both the magnitude and sign of the estimates vary considerably depending on the choice of controls one includes. These headline results likewise disappear with correction for nonstationary time series.

The clear takeaway is that panel IV estimation that assumes iid error terms and ignores the temporal sequencing of observations runs a serious risk of spurious regressions given the high likelihood of co-trending variables. This manifests in both mistaken inference and parameter estimates biased in the direction of the reverse causality that motivated the use of an IV estimator in the first place.

Table 3: First-differenced 2SLS coefficients of food aid on conflict

	Dependent Variable: Dummy for war in year t - dummy for war in year (t-1)						
	(1) Any War	(2) Any War	(3) Any War	(4) Any War	(5) Any War	(6) Intrastate	(7) Interstate
ΔAid_t	-0.00801	-0.0115	-0.00858	-0.00727	-0.0659	-0.0779	-0.0163
	(0.013)	(0.024)	(0.015)	(0.009)	(0.493)	(0.582)	(0.113)
Observations	4034	4034	4034	4034	3964	3964	3964
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-year linear trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes
US real per capita GDP x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
US democratic president x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Oil price x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Monthly recipient temperature. and precipitation	No	No	Yes	Yes	Yes	Yes	Yes
Monthly weather x avg. prob. of any US food aid	No	No	Yes	Yes	Yes	Yes	Yes
Avg. US military aid x year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. US economic aid (net of food aid) x year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. recipient cereal imports x year FE	No	No	No	No	Yes	Yes	Yes
Avg. recipient cereal production x year FE	No	No	No	No	Yes	Yes	Yes
KP weak F-Stat	.44	.23	.36	1.38	.02	.02	.02

Notes: This table replicates the 2SLS estimates from Table 2 in NQ, using the same set of controls as NQ and clustering at the country level as in NQ and HI. The change from NQ involves replacing the level values of food aid, conflict and wheat production with first differenced values. For example, ΔAid_t is the quantity of wheat food aid delivered (in metric tons, MT) in year t minus the quantity delivered in year t-1. The instrument for the 2SLS estimate of the effect of ΔAid_t is $\Delta wheat_{t-1}$, where $\Delta wheat_{t-1}$ is the quantity of wheat produced in the US (in 100,000 MT) in year t-1 minus the quantity of wheat produced in year t-2.

Table 4: First-differenced 2SLS coefficients of GDP growth on conflict

	Dependent Variable: Dummy for war in year t - dummy for war in year (t-1)						
	(1) Any War	(2) Any War	(3) Any War	(4) Any War	(5) Any War	(6) Intrastate	(7) Interstate
$\Delta Ln(GDP)_t$	-13.03	27.30	-663.5	-5.494	5.036	-4.358	5.631
	(52.064)	(179.978)	(1.07e+05)	(8.915)	(5.492)	(5.085)	(5.105)
Observations	4000	4000	4000	4000	3930	3930	3930
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-year linear trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes
US real per capita GDP x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
US democratic president x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Oil price x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Monthly recipient temperature. and precipitation	No	No	Yes	Yes	Yes	Yes	Yes
Monthly weather x avg. prob. of any US food aid	No	No	Yes	Yes	Yes	Yes	Yes
Avg. US military aid x year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. US economic aid (net of food aid) x year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. recipient cereal imports x year FE	No	No	No	No	Yes	Yes	Yes
Avg. recipient cereal production x year FE	No	No	No	No	Yes	Yes	Yes
KP weak F-Stat	.07	.02	.00004	.55	1.44	1.44	1.44

Notes: This table replicates the 2SLS estimates of gdp growth on conflict, using the same set of controls as in Table 2 of NQ and clustering at the country level as in NQ and HI. The change from HI involves replacing the level values of interest rates and conflict with first differenced values. For example, ΔR_t is global interest rate in year t minus the interest rate in year t-1. The instrument for the 2SLS estimate of the effect of $\Delta Ln(GDP)_t$ is ΔR_{t-1}.

4. Panel IV Methods With Interacted Instruments

The possibility remains, however, that a true causal relation really underlies the observed correlations reported in HI, NQ, and other papers that rely on identification by panel IV methods. HI and NQ – and many other authors – rely on a shift-share/Bartik or similar interacted instruments to try to identify a causal effect of an endogenous explanatory variable of interest. In this section we show that although interacting the Z_t time series instrumental variable with another variable that varies only in the cross-section buys some added flexibility in accommodating time trends, the interaction does not ameliorate the spurious regressions problem.

In practice, the interacted instrument strategy is implemented by estimating variants of the two following equations:

$$Conflict_{it} = \gamma^{int} Z_t * D_i + \boldsymbol{Controls_{irt}}\Gamma^{int} + \theta_i^{int} + \rho_{irt}^{int} + \mu_{it}^{int} \quad (32)$$

$$X_{it} = \pi^{int} Z_t * D_i + \boldsymbol{Controls_{irt}}\Gamma^{\text{int}} + \Theta_i^{int} + \text{P}_{irt}^{int} + \eta_{it}^{int} \quad (33)$$

Such a strategy requires selecting a variable, D_i, that varies in the cross-section to interact with the exogenous time series variable. NQ use the regularity of food aid receipts, defined as the proportion of the 36 years in their sample data in which country *i* received any food aid from the US. HI use three different variables for each country *i*: whether it used a fixed exchange rate, a measure of the openness of the country's capital account to financial flows, and a measure of ethnolinguistic and religious fractionalization to measure within-country sociocultural diversity.

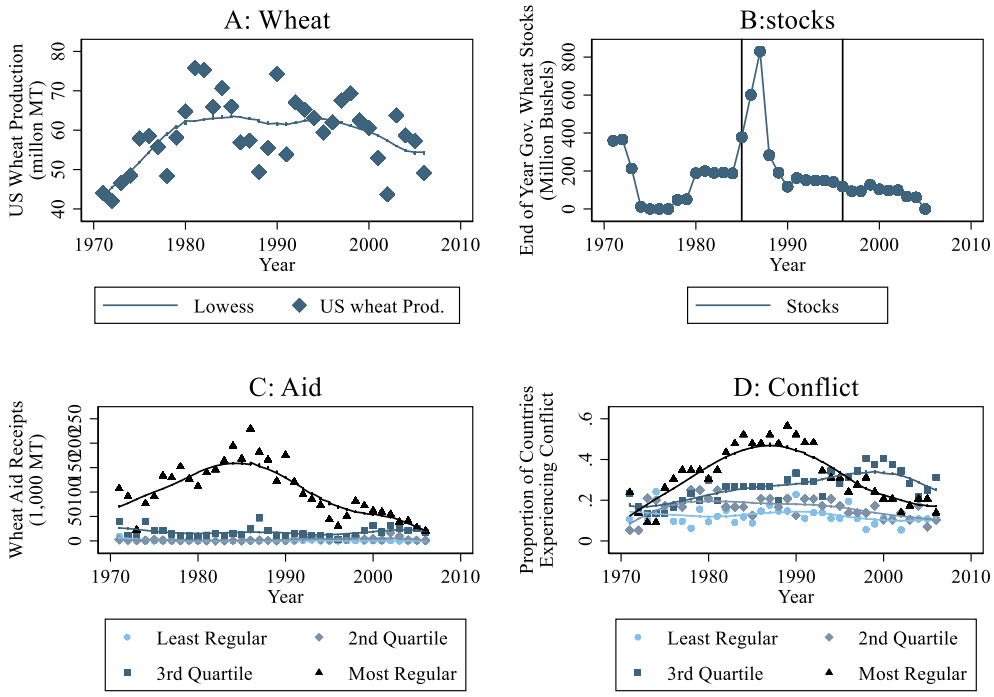
Relative to the uninteracted equations (Equations 24 and 25), this specification introduces two important changes. First, the interaction allows for the possibility of differential exposure to the effect of interest, as the transmission of the time series innovation in Z_t is mediated by the cross-sectional exposure variable, D_i. When D_i is a dummy variable, like an indicator for a country operating a fixed exchange rate regime, this functions like a standard DiD estimator. When D_i is continuous this resembles a dose-response estimator.

Second, the instrument is interacted in both the reduced form equation (32) and the first stage equation (33). This allows for more flexible, nonparametric accommodation of unknown common trends, where ρ_{irt}^{int} and P_{irt}^{int} are year fixed effects instead of the linear time trend *t* included in equations 24 and 25 with the uninteracted instrument.

As we saw in the simple model in Section I, this strategy only allows the researcher to control for time trends that are *common* to the countries of the various types described by the continuum of variation in the variable *D*. Although the strategy can avoid the need to parameterize unobserved trends, the requirement that the shift-variable not be correlated with heterogeneity in trends is a stronger caveat than it may seem. In the context of the NQ and HI cases, if the problematic trend in conflict only appears (or appears more strongly) in countries that both experience conflict and more regularly received food aid or exhibit less ethnolinguistic fractionalization, then adding the flexible time trend does not remove

the endogeneity. Below we describe how this problem arises and can be diagnosed in the NQ case.[27]

Figure 8: Time trends in the NQ variables



Notes: All variables are taken from NQ dataset, except for government wheat stocks, which is taken from the Farm Service Agency and National Agricultural Statistics Service, USDA (2006). Includes 127 countries from 1971-2006. Trend lines in panels A, C, and D are estimated by lowess.

We see how the interaction strategy arises in the NQ setup by plotting the temporal variation of the key NQ variables and separate these trends on the same dimension as the interaction strategy. Figures 8a and 8b show the first stage intuition of the policy mechanism that motivates the NQ identification strategy. When lagged US wheat output is high, US government grain purchases lead to accumulation of stocks that get distributed the next year as food aid. Figure 8c visualizes the NQ shift-share identification strategy,

[27] Jaeger et al. (2020) offer a similar critique of Kearney and Levine (2015), demonstrating the fragility of the identifying assumption that trends across groups are identical, and explaining why the interacted instrument fails to resolve the endogeneity problem that confounds causal interpretation of the observed partial correlation.

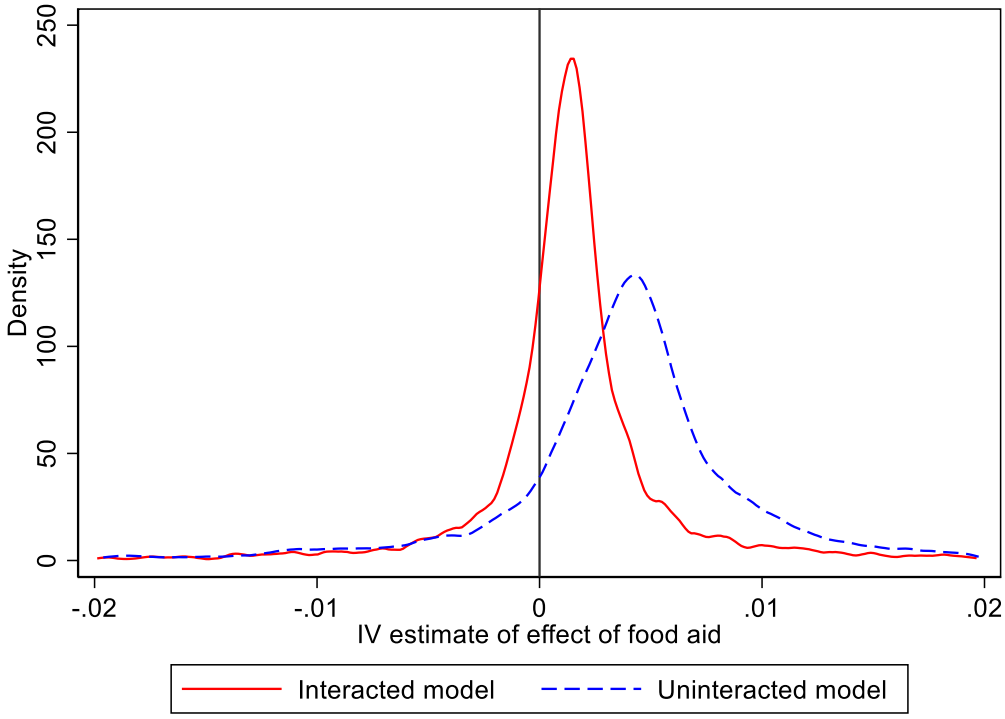
showing that food aid flows to the most regular quartile of recipient countries indeed tracks lagged US wheat output and US government wheat stocks reasonably well. The inverse-U trend that clearly appears among the most frequent aid recipients is not present among the infrequent recipients. Replicating this exercise for the conflict variable in Figure 8d reveals a similar pattern in conflict. Regular food aid recipients have a strong inverse-U trend in conflict that does not appear among the least frequent recipients.

We now introduce the interaction term into the Monte Carlo simulation setup used in section 3 to show that the bias and inference issues that arise from spurious trends in the uninteracted case remain with the shift-share interacted IV method that incorporates period fixed effects. Figure 9 shows the estimated 2SLS coefficients of food aid on conflict from 1,000 simulations using the same spurious instrument with a random walk as before.[28] Controlling flexibly for underlying, common time trends does not mean that irrelevant instruments will be equally likely to return positive coefficients as negative ones, because the distribution of IV coefficient estimates in the interacted case remains centered above zero with only 23.8% of simulations returning a negative coefficient. The estimates are merely rescaled, not moved by the interaction variable. Just as in the uninteracted case, using a spurious, non-stationary time series variable as an instrument in expectation returns a positive and statistically significant estimated effect of food aid on conflict. Importantly, this effect is identified only via a common cyclical trend in both aid and conflict that is not shared by both regular and irregular recipients of aid.

As before, a causal effect of aid on conflict is one possible explanation for this association, but it could equally be spurious. Using a panel IV approach in no way ensures causality because of the spurious regression problem. Statistical power differs across the two samples, with a tighter distribution of coefficients in the interacted case than the uninteracted case can occur.

[28] Model 3 in Appendix B replicates these findings in the fully simulated context with the inclusion of a shift-share type of instrument and time period fixed effects. We show that even when endogeneity in the first stage is minimized, if the cross-sectional component of the interacted shift-share instrument is endogenous, this generates bias in the resulting IV coefficient estimates.

Figure 9: Simulated distribution of 2SLS estimates using shift-share instrument



Notes: Distributions from simulating a fake instrument Z_t 1,000 times. Outcome variables and controls are taken from the NQ replication dataset and the NQ baseline specification. Interacted model is the 2SLS coefficient estimated from the first stage and reduced form equations described by equations (32) and (33) and Uninteracted model is the 2SLS coefficients estimated from equations (26) and (27)

To understand how interactions affect the reliability of weak instrument tests, we test in our simulations how well weak instruments tests correctly categorize our known-to-be-irrelevant instruments. We can compare rejection rates of tests for weak instrument F-statistics being greater than 10 in our simulations of irrelevant instruments. We find that only 2.2% of irrelevant instruments pass this test in uninteracted models, suggesting that this test does fairly well in identifying weak instruments. In the interacted models, the F=10 threshold is exceeded in 3.49% of cases. It therefore seems that F-tests are reasonably well powered to reject these placebo instruments. However, if we consider the cases where F-statistics are above 10 in either the interacted or the uninteracted model, we find that the F-statistic for one or both of these regressions passes the threshold of 10 in 5.43% of

simulations. The risk of specification searching then becomes relevant. Because F-tests are weakly correlated across models, allowing authors to report an F-stat below 10 for one specification as long as the other passes 10 reduces the power of the weak instrument test. For comparison, NQ's F-statistic for their uninteracted specification is 3.35 and 12.1 for the interacted specification. We describe these results further in Appendix D, and also show that the distribution of IV coefficients estimated by interactions with irrelevant instruments is more biased among the instruments that pass weak instrument tests in interacted models.

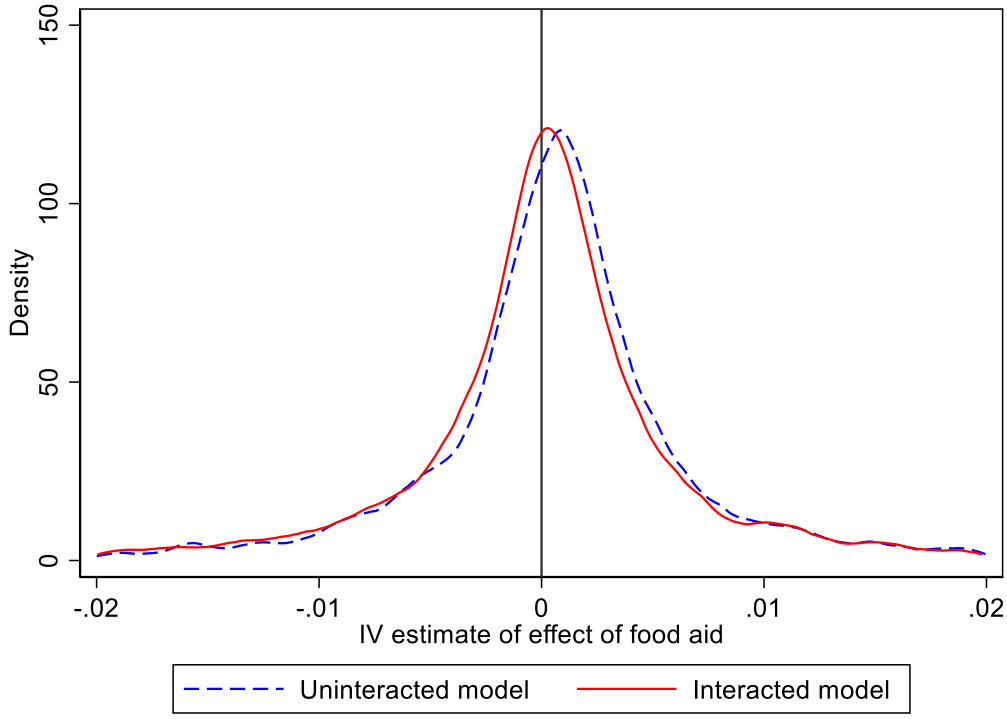
We showed via simulations of the uninteracted models that first differencing the dependent variable, the instrument, and the endogenous regressor corrects for a known-I(1) instrument. We re-estimate the interacted and uninteracted models in 1,000 simulations now using the first differenced variables.[29] Figure 10 shows the distribution of estimated 2SLS coefficients of food aid on conflict. First differencing all relevant variables eliminates the risk that the estimated coefficients will be distributed around a non-zero value for both the interacted and uninteracted specification, suggesting that this check reduces the risk that an irrelevant instrument would return an association consistently more likely to have one sign than the other.

[29] For the interacted models, the specification is:

$\Delta Conflict_{it} = \gamma^{intdiff} \Delta Z_t * D_i + t_t^{intdiff} + \mu_{it}^{intdiff}$ for the reduced form and

$\Delta X_{it} = \pi^{intdiff} \Delta Z_t * D_i + T_t^{intdiff} + \eta_{it}^{intdiff}$ for the first stage.

Figure 10: 2SLS estimates using interacted instrument and first-differenced variables



Notes: Distributions from simulating a fake instrument Z_t 1,000 times. Outcome variables and controls are taken from the NQ replication dataset and the NQ baseline specification.

The take-away message of this section is simple: the interacted instrument does not solve the spurious regressions problem that easily arises in the time series component of a panel, regardless of whether it satisfies the standard relevance and exclusion criteria for instrumental variables. Interacting the instrument that varies in time series with another variable that varies in cross-section merely rescales the ILS/2SLS estimates from the uninteracted regression specifications. If the weights are endogenous to the outcome and the unobserved trend, the interacted specification could have an even greater risk of being centered around a value other than the true causal value. Because F-statistics are not perfectly correlated across interaction specifications with different possible interactions, a rule which considers an instrument valid if it passes a weak IV test for at least one specification reduces power of these tests and may increase bias.

One corrects the bias only by correcting for the spurious correlation arising in the time series. In the limiting case of weak endogeneity and a strong first stage, this can be done for the using a heteroskedasticity-and-autocorrelation consistent estimator, such as Newey-West to make correct inference in the first stage. But in the presence of significant endogeneity, bias arises and correcting the standard errors will not suffice. We find that first differencing I(1) variables works well where the instrument is known to follow an I(1) random walk process. Then the interaction adds no statistical power to the uninteracted regressions, while it retains the other advantages of the interacted instrument design: more flexible accommodation of common trends and more nuanced interpretation of the coefficient estimates in a manner consistent with difference-in-difference or dose-response estimators.

5. Diagnostic Steps For Panel IV Estimation

The preceding cautions notwithstanding, in some cases panel IV estimation may work for causal identification of a relationship of interest. Authors and consumers of research need to consider when they need to apply corrections for non-stationarity or serial correlation and how to decide which findings to prefer when results are sensitive to trend or stationary corrections. Several steps can help address these concerns.

Authors should carefully visually inspect their data, presenting primary variables on the dimensions of natural sequencing patterns in the data generating process, either on the time dimension as we show here for time series and panel applications, or on maps when autocorrelation is likely to be spatial as advocated by Kelly (2020). Autocorrelation should be reported for primary variables including outcomes, instruments, and endogenous variables of policy interest. High autocorrelation in any of these should trigger concern. Tests of trend and difference stationarity such as Augmented Dickey Fuller tests for single time series or Fisher-type and Hadri tests in panel data can help identify the most serious deviations from stationarity. But these tests may have low power in many applications and may fail to reject stationarity of some panels if at least some panels are stationary. Tests for serial correlation such as those proposed by Born and Breitung (2016) or Wooldridge

(2002) can test for serial correlation more generally. Addressing serial correlation satisfactorily can resolve the mistaken inference problem intrinsic to spurious regressions.

The greater concern is when the explanatory variable of interest is strongly endogenous. Particularly when using specifications that interact time series variables with cross-sectional ones, authors should defend their belief in exogeneity of shocks (Borusyak et al., 2020) or shares (Goldsmith-Pinkham et al., 2020) and should carefully interpret differences between OLS and IV results with a view toward whether differences could be explained by deviations from these assumptions. Authors should report multiple specifications for deterministic trends with corresponding weak instrument tests and Anderson-Rubin tests for each specification. Linear trend controls or time period fixed effects may not justify exclusion restrictions if deterministic trends are non-linear or arise from the persistent sum of past shocks. When policy changes allow for pre-period tests or structural breaks, these tests can help identify sources of non-parallel trends.

Finally, randomization inference, randomized placebo tests, and simulations of models with known data generating process can be very helpful in diagnosing the role and magnitude of primary threats to both inference and identification. If signs of mistaken inference arise from these tests, we recommend that authors report specifications with all variables first differenced and interpret differences in conclusions. For example, more careful theory may be required to justify the adjustment dynamics that explain why effects appear in long term trends that are not apparent in year-to-year changes.

No one test or specification can diagnose or solve all of the possible threats to identification in panel IV studies, and other proposals may be appropriate for other contexts. In shift-share IV specifications, residualization methods proposed by Goldsmith-Pinkham et al. (2020) or Borusyak et al. (2020) can be used to validate the plausibility of identification from exogenous shocks or exogenous shares. The methods proposed by Adão et al. (2019) can help address concerns about inference in Bartik/shift-share instruments when cross-sectional residuals are correlated across units with similar shares. When assignment rules to exogenous shocks are endogenous, but known, the assignment rule can be used to de-bias the endogeneity as proposed by Borusyak and Hull (2021). Jaeger et al's

(2018) proposal to include the lagged shift share to instrument for the lagged endogenous regressor can help address dynamic adjustment may mitigate bias arising from the lagged adjustment to a stable endogenous regressor of interest in a similar manner to the differencing primary variables as we have advocated. Young (2019) offers useful diagnostics for cases where single clusters are highly leveraged, while Kelly et al (2020) does likewise for cases of high spatial autocorrelation. Each of these show promise for both diagnosing issues and recovering valid instruments in their special cases.

6. Conclusions

In this paper, we show that a panel data IV estimation strategy that has become popular among researchers seeking to identify the causes of conflict and other key outcomes may be subject to heretofore unrecognized inferential errors and bias. The most likely source of error arises from spurious regressions if the time series properties of the panel variables render the regression errors non-iid. Interacted (e.g., Bartik/shift-share) instruments and year fixed effects do not resolve that problem. Indeed, choosing endogenous weights in the shift-share instrument without justifying an exclusion restriction for the weighting variable can inadvertently allow authors to select interactions that satisfy a first stage weak instruments test by reweighting finite sample bias, without actually removing the bias.

Much like Bazzi and Clemens (2013), we offer a caution about instrument validity and strength in panel data IV estimation. We show that simple diagnostics call into question whether key variables in the specifications reported by prominent papers in the causes of conflict literature can be reasonably assumed stationary, and that inference is therefore flawed and finite sample bias issues arise when the iid assumption is violated. Familiar corrections to standard errors for serial correlation will not suffice in the presence of an endogenous regressor. First differencing to render the instrument, explanatory and outcome variables stationary appears a reasonably promising way to address the spurious regression problem in panel IV estimation.

REFERENCES

Adao, R., Kolesár, M. and Morales, E., 2019. Shift-share designs: Theory and inference. *Quarterly Journal of Economics*, *134*(4), pp.1949-2010.

Barrett, Christopher B. (1998). "Food Aid: Is It Development Assistance, Trade Promotion, Both or Neither?" *American Journal of Agricultural Economics* 80(3): 566-571.

Barrett, Christopher B. and Daniel G. Maxwell (2005). *Food Aid After Fifty Years: Recasting Its Role.* New York: Routledge.

Bartik, Timothy J. (1991). *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Bazzi, Samuel and Michael A. Clemens (2013). "Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth." *American Economic Journal: Macroeconomics* 5(2): 152-186.

Bertrand, Marianne, Duflo, Esther. and Mullainathan, Sendhil., 2004. How much should we trust differences-in-differences estimates?. *The Quarterly journal of economics*, *119*(1): 249-275.

Born, Benjamin, and Jörg Breitung. "Testing for serial correlation in fixed-effects panel data models." *Econometric Reviews* 35, no. 7 (2016): 1290-1316.

Blattman, Christopher, and Edward Miguel (2010). "Civil war." *Journal of Economic Literature* 48(1): 3-57.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2018). "Quasi-experimental shift-share research designs." National Bureau of Economic Research Working paper 24997.

Chu, Chi-Yang, Daniel J. Henderson, and Le Wang (2017). "The Robust Relationship Between US Food Aid and Civil Conflict." *Journal of Applied Econometrics* 32(5): 1027-32.

Enders, Walter (2008). *Applied Econometric Time Series.* John Wiley and Sons.

Ernst, Philip A., Larry A. Shepp, and Abraham J. Wyner (2017). "Yule's "Nonsense Correlation" Solved!" *Annals of Statistics* 45(4): 1789-1809.

Farm Service Agency and National Agricultural Statistics Service, USDA (2006). "Appendix table 9--Wheat: Farm prices, support prices, and ending stocks, 1955/56-

2005/06." Accessed 14 May 2015. www.ers.usda.gov/webdocs/DataFiles/

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020), "Bartik Instruments: What, When, Why, and How", *American Economic Review* 110(8): 2586-2624.

Granger, Clive W.J., and Paul Newbold (1974). "Spurious regressions in econometrics." *Journal of Econometrics* 2(2): 111-120.

Hull, Peter, and Masami Imai (2013). "Economic shocks and civil conflict: Evidence from foreign interest rate movements." *Journal of Development Economics* 103: 77-89.

International Federation of the Phonographic Industry (IFPI). "Recording Industry in Numbers." Accessed 26 Aug 2015. https://musicbusinessresearch.wordpress.com/ 2010/03/29/the-recession-in-the-music- industry-a-cause-analysis/.

International Monetary Fund. "Interest Rates selected indicators." International Financial Statistics, (IFS), https://data.imf.org/regular.aspx?key=61545855. Accessed April 30, 2021.

Jaeger, David A., Theodore J. Joyce, and Robert Kaestner (2019). "Tweet Sixteen and Pregnant: Missing Links in the Causal Chain from Reality TV to Fertility." *International Journal for Re-Views in Empirical Economics* 3.

Jaeger, David A., Theodore J. Joyce, and Robert Kaestner (2020). "A Cautionary Tale of Evaluating Identifying Assumptions: Did Reality TV Really Cause A Decline in Teenage Childbearing?" *Journal of Business and Economic Statistics* 38(2): 317-326.

Jaeger, David A., Joakim Ruist, and Jan Stuhler (2018). "Shift-share instruments and the impact of immigration." NBER working paper 24285.

Kearney, Melissa S. and Phillip B. Levine (2015). "Media Influences on Social Outcomes: The Impact of MTV's *16 and Pregnant* on Teen Childbearing." *American Economic Review* 105(12): 3597-3632.

Keller, Wolfgang (1998). "Are international R&D spillovers trade-related?: Analyzing spillovers among randomly matched trade partners." *European Economic Review*, 42(8): 1469-1481.

Kelly, Morgan (2019). "The standard errors of persistence." Unpublished Manuscript.

Mikusheva, A. (2007). Uniform inference in autoregressive models. Econometrica, 75(5),

1411-1452.

Nickell, Stephen (1981). "Biases in dynamic models with fixed effects." *Econometrica* 49(6): 1417-1426.

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

Phillips, Peter C.B. (1986). "Understanding spurious regressions in econometrics," *Journal of Econometrics* 33(3): 311-340.

Phillips, Peter C.B., and Bruce E. Hansen (1990). "Statistical Inference in Instrumental Variables Regression with I(1) Processes." *Review of Economic Studies* 57(1): 99–125.

Phillips, Peter C.B. (1998). "New tools for understanding spurious regressions." *Econometrica* 66(6): 1299-1325.

Shambaugh, Jay (2004). "The effects of fixed exchange rates on monetary policy." *Quarterly Journal of Economics* 119(1): 301-352.

Slutzky, Eugen (1937). "The summation of random causes as the source of cyclic processes." *Econometrica*: 105-146.

USAID (2014). "(Re)Assessing The Relationship Between Food Aid and Armed Conflict." USAID Technical Brief.

Willis, Brandon and Doug O'Brien. "Summary and Evolution of U.S. Farm Bill Commodity Titles." National Agriculture Law Center. Accessed 26 January 2015. http://nationalaglawcenter.org/farmbills/commodity/

World Bank. "Real Interest Rate (%) World Development Indicators, The World Bank Group, https://data.worldbank.org/indicator/FR.INR.RINR. Accessed May 2, 2018.

Wooldridge, J.M., 2002. *Econometric analysis of cross section and panel data*. MIT Press.

Young, Alwyn (2018). "Consistency without inference: Instrumental variables in practical application." Unpublished manuscript. London School of Economics and Political Science.

Yule, G. Udny (1926) "Why do we sometimes get nonsense-correlations between Time-Series?--a study in sampling and the nature of time-series." *Journal of the Royal Statistical Society* 89(1): 1-63.

Zürcher, Christoph (2017). "What do we (not) know about development aid and violence? A systematic review." *World Development* 98: 506-522.

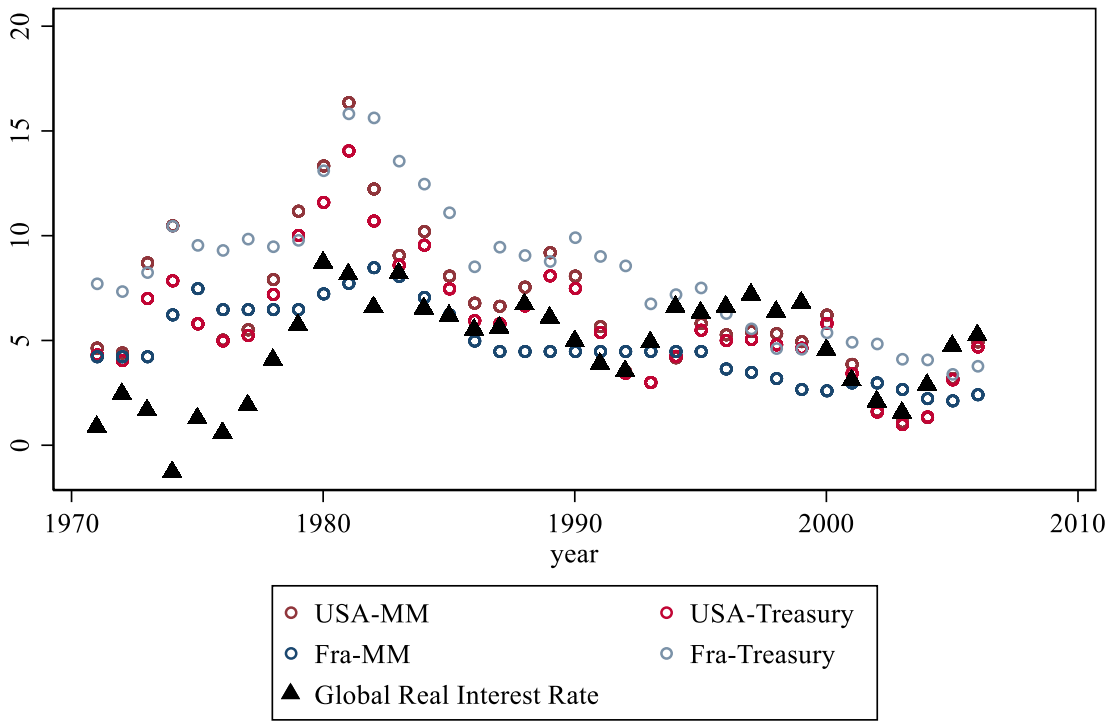
Appendix To Spurious Regressions and Panel IV Estimation: Revisiting the Causes of Conflict (For Online Publication Only)

By PAUL CHRISTIAN AND CHRISTOPHER B. BARRETT

Appendix A: Estimating HI specifications with multiple base country interest rates

In the main body of the paper, we use the average real interest rate (R_t) as an instrument for GDP growth. The average real interest rate does not vary across countries over time. HI in the primary specification use the base interest rate for the country and interest rate suggested by Shambaugh (2004) as the most relevant for a country i with an open economy pegged to or influenced by the base country. This means that the instrument Z_t varies in the cross-section, so that the base interest rate *Base* $R_{t-1,i}$ is indexed by country *i*. In practice, most countries are pegged the same small number of countries. In Shambaugh (2004), more than half of countries' base country is the United States, and the US and France together account for the base countries for more than 69% of countries. When considering countries that appear in the NQ dataset, the US and France are the base countries for fully 81.6% of countries in the panel. The interest rates among the base countries also move on the same trends (Figure A1). In the most important countries, interest rates followed the dominant trends of global real interests, higher in the 1980's and 1990's and lower in the 1970's and 2000's.

Figure A1: Trends in interest rates in France, USA, and global real interest rates



Notes: Data from IMF for USA and France Base country rates as described in HI and Shambaugh (2004), and from World Development Indicators for Global Real Interest Rates

Given that interest rates across countries are likely to be highly correlated due to a non-arbitrage condition, it is unlikely that using base country interest rates yields different results from using average real interest rates. If base country interest rates are determined by a model like $b_i R_t = b_i(R_t + e_t)$, then regressing conflict on base interest rates will identify the same spurious correlation from the persistence in R_t that we would find in $b_i R_t$. However, it may be the case that cross sectional variation in base country interest leads to meaningfully different results. To be sure that we are accurately addressing the core issue in HI, we replicate the approach using base interest rates as described by HI.

To replicate the HI main model on the data available from NQ's replication file, we estimate:

$$\Delta\text{GDPt}_{(it)} = \theta_i + \theta_i^{trend} t + \beta_1 R^b_{i(t-1)} + \nu_{(it)} \quad \text{(A1)}$$

$$conflict_{(it)} = \gamma_i + \gamma_i^{trend} t + \delta \hat{y}_{(it)} + \epsilon_{(it)} \quad \text{(A2)}$$

To assess whether we would have found the same result if we had used a single global real interest rate, we estimate the following two modified specifications, replacing $R^{b}_{i(t-1)}$ with the global average real interest rate from the World Development Indicators database as follows:

$$\Delta\text{GDPt}_{(it)} = \theta_i + \theta_i^{trend} t + \beta_1 R_t + \nu_{(it)} \quad \text{(A3)}$$

$$conflict_{(it)} = \gamma_i + \gamma_i^{trend} t + \delta \hat{y}_{(it)} + \epsilon_{(it)} \quad \text{(A4)}$$

One implication of using base interest rates rather than a global interest rate that does not vary across countries is that Shambaugh (2004) does not propose a base country for every country in the NQ dataset, so using the base interest rate specification limits the number of observations we can use. We can match 59 out of the 127 countries from the NQ dataset to a base country. To assess whether the resulting sample restriction influences results, we estimate equations A3 and A4 with the entire NQ sample, as using R_t rather than $R^{b}_{i(t-1)}$ allows us to avoid dropping countries that we cannot match to a base country b.

Finally, NQ and HI differ in how they estimate trends in uninteracted models. HI interact trends with a fixed effect for every country as shown in equations A1-A4. In contrast, NQ estimate a common trend for all countries as in the same region. We estimate this specification as well to compare to the role of estimating country specific trends when the instrument varies by country.

The first stage results are shown in table A1. Using the countries that match between the NQ data and Shambaugh, we find nearly identical results as HI. The first stage coefficient (column 1, table A1) is -0.402, compared to HI's estimated -0.302. Changing to a fixed real interest rate across countries changes the coefficient to -0.338, even closer to the estimate reported by HI. Using the fixed interest rate in the full NQ sample (column 3) or using the NQ trend specification (column 4) have negligible influence on the first stage.

Table A1: Replicating HI first stage using base or average real interest rates as IV

VARIABLES	(1) ΔGDP_t	(2) ΔGDP_t	(3) ΔGDP_t	(4) ΔGDP_t
Base $R_{t-1,i}$	-0.402***			
	(0.0677)			
Real Interest Rate$_t$		-0.338***	-0.342***	-0.368***
		(0.0970)	(0.0749)	(0.0716)
t*country	Yes	Yes	Yes	No
t*wb_region	No	No	No	Yes
Observations	1,981	1,981	4,087	4,087
R-squared	0.113	0.105	0.124	0.071

Notes: Robust standard errors in parentheses, clustered at country level. *** p<0.01, ** p<0.05, * p<0.1
GDP is taken from NQ dataset, Base rates rates are merged to the NQ dataset from IMF, (2021), and Real Interest Rates is Merged from World Bank (2018).

The 2SLS results are shown in Table A2. Again, using the HI Base rate model in the NQ data that can be matched yields nearly identical results to HI's specification. We find an estimated association between GDP growth and conflict of -2.70093 with a standard error of 1.1 compared to HI's reported -2.40 and standard error of 1.08. When using the fixed interest rates in the full NQ sample (column 3) or the NQ trend specification (column 4), the results are again nearly identical to the main specification reported by HI. The K-P F statistics estimated across models range from 16 to 24, slightly lower than the 35.8 reported by HI for their base model, but still well above the usual rule of thumb of 10.

Table A2: Replicating HI 2SLS using base or average real interest rates as IV

VARIABLES	(1) Any_war$_t$ (IV is Base $R_{t-1,i}$)	(2) Any_war$_t$ (IV is R_t)	(3) Any_war$_t$ (IV is R_t)	(4) Any_war$_t$ (IV is R_t)
ΔGDP_t	-2.70093**	-1.79315	-3.05009**	-2.97860***
	(1.09998)	(1.42018)	(1.23020)	(1.07542)
t*country	Yes	Yes	Yes	No
t*wb_region	No	No	No	Yes
Observations	1,981	1,981s	4,016	4,016
K-P rk F Stat.	20.909	24.21	16.027	24.385
R-squared	0.50422	0.59050	0.31139	0.20827

Notes: Robust standard errors in parentheses, clustered at country level. *** p<0.01, ** p<0.05, * p<0.1
GDP is taken from NQ dataset, Base rates rates are merged to the NQ dataset from IMF, (2021), and Real Interest Rates is Merged from World Bank (2018).

Appendix B: Simulation results under varying parameterizations

In this Appendix we generalize from the example presented in the main paper, using a fully controlled system of equations of known parameterization. In each model, the true causal parameter of interest, β from equation (1), equals zero. Among simulation models, we vary the degree of serial correlation and the relative strength of the instrument and of the endogeneity in the first stage, as reflected in the coefficients χ and α, respectively, from equation (2). Within each model, we simulate under varying degrees of persistence of the random innovations in the time series, ρ, from the set $\rho = \{0.0, 0.1, 0.5, 0.6, 0.9, 1.0\}$, thus ranging from iid through a fully I(1) time series.

We start, in model 1, with the case of a strong instrument and weak endogeneity. All the parameter estimates are unbiased and the sampling distribution of is reasonably behaved when $\rho=0$ but their sampling distributions become increasingly diffuse as ρ increases. This essentially replicates the canonical spurious correlation result from the time series literature.

In model 2, the endogeneity is made prominent. In the first stage and reduced form equations, the coefficient estimates are unbiased but suffer from mistaken inference as ρ increases. But because the first stage and reduced form coefficient estimates are strongly, spuriously correlated pronounced bias emerges in the IV coefficient estimates, in the same direction of the reverse causality the IV is meant to resolve. Unlike the sampling distributions of the first stage and reduced form estimates, the sampling distribution of the IV estimates becomes more rather than less concentrated as ρ increases, resulting in firmer erroneous rejection of the null. This is the finite sample bias problem previously unrecognized in panel IV estimators.

Model 3 combines the features of the first two models. In this model, a first stage is strong enough to dominate when serial autocorrelation is low, but as we increase that parameter, realizations of $\widehat{cov}(\tau_t, Z_t)$ become large enough to swamp the true first stage. The relevant finding from this result is that the bias in the IV is not necessarily constant over degrees of persistence as it was in the limit cases of the first two models. In this

model, the role of correcting the autocorrelation in this case by first differencing appears in multiple places. Now first differencing reduces the risk across iterations of large spurious coefficients in both the reduced form and the first stage, and centers the IV distribution closer to the true causal value, with these effects more important for more persistent systems.

B1. Model 1: Strong instrument and minimal endogeneity

Model 1 is just a parameterized version of the model laid out in equations (15) and (16) in section 2. Conflict, c_{it}, is a random process described by a shared time effect scaled by country effects. An endogenous variable X_{it} is a function of contemporaneous conflict and an exogenous time series variable, Z_t. In model 1, we assume there is no causal effect of X_{it} on conflict, but we are interested in how the estimated first stage, reduced form and 2SLS coefficients estimated in a finite sample are affected by persistence in the annual shocks to conflict and the instrument that drive time series variation in these variables. In model 1, the instrument is relevant and the simultaneity the IV is meant to overcome is rather weak (0.005), nearing the canonical time series spurious regressions case of truly independent time series.

The data generating process we simulate is characterized by:

$$c_{it} = 1.5\psi_i\tau_t + \sigma_{it} \quad \text{(B1)}$$

$$X_{it} = .005c_{it} + Z_t + \eta_{it} \quad \text{(B2)}$$

$$\tau_t = 100 + \rho(\tau_{t-1} - 100) + s_t \quad \text{(B3)}$$

$$Z_t = 100 + \rho(Z_{t-1} - 100) + q_t \quad \text{(B4)}$$

$$\psi_i \sim U(0,10) \quad \text{(B5)}$$

$$\sigma_{it} \sim N(0,10) \quad \text{(B6)}$$

$$\eta_{it} \sim N(0,10) \quad \text{(B7)}$$

$$s_t \sim N(0,10) \quad \text{(B8)}$$

$$q_t \sim N(0,10) \quad \text{(B9)}$$

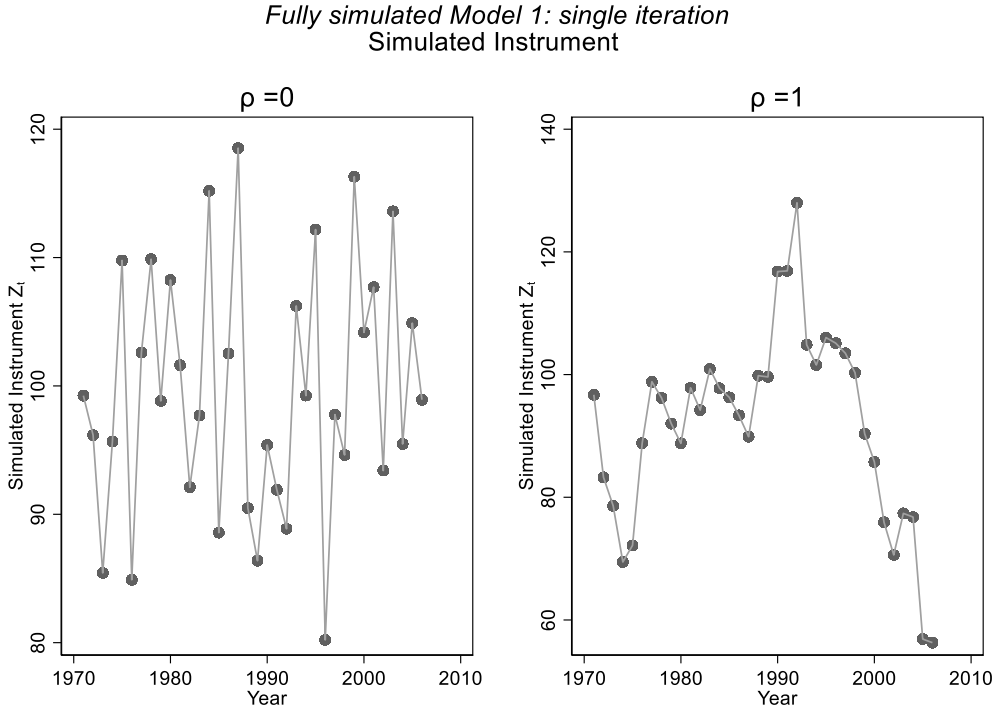
In terms of the conflict and food aid example, this captures a system in which countries i in year t share a risk of conflict affected by the unobservable variable τ_t. The

influence of this time effect varies by country as it is scaled by a country effect ψ_i, capturing the idea that "bad" years for conflict are particularly likely to increase conflict in some countries more than others. Aid decisions are endogenous to contemporaneous conflict, creating the concern that estimating a relationship between c_{it} and X_{it} by OLS would be biased by simultaneity. Because Z_t appears in the model for X_{it}, but not in the model for c_{it}, it is a theoretically good candidate for an instrument.

Our goal with this simple model is to show persistence of the instrument and the time series dimension of the unobservable influences on conflict affect inference – in model 2, bias – in estimated coefficients by adjusting ρ. When $\rho = 0$, both the instrument and the observable dimensions of conflict are white noise around a constant, and when $\rho = 1$, the instrument and the unobservable determinants of conflict follow a random walk around the starting constant.

Before we estimate the primary equations of interest, we can plot the variables of interest over time in a single iteration of the simulated dataset. In each iteration of the simulation, we estimate this system of equations with N= 126, and T= 35, chosen to match the NQ dataset. The figure below shows the instrument Z_t. The left plot is a case where the shocks to the instrument do not persist ($\rho = 0$), and the time series looks like random noise around the mean value. In the right plot, the instrument follows a random walk from the constant ($\rho = 1$), showing much smoother transitions from year to year and more variation over longer time horizons.

Figure B1: Fully simulated datasets with varying time-series persistence

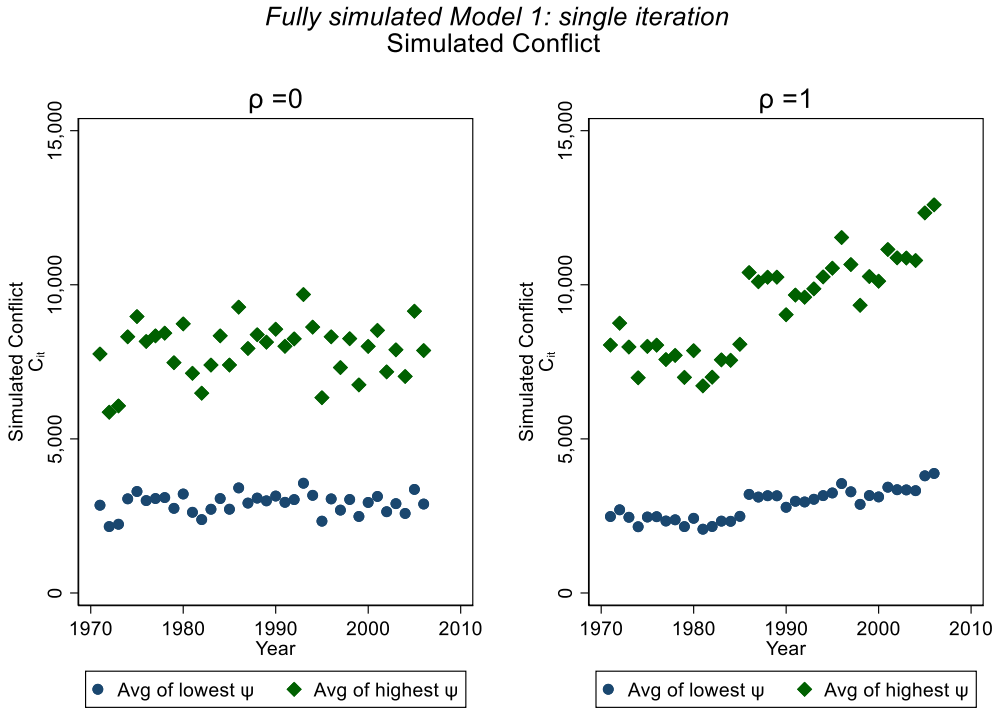


Notes: Data in this figure are from a fully simulated dataset with N=125, T=36 simulating random draws for equations B1-B9.

The next figure shows simulated conflict for a series with white noise errors around a constant ($\rho = 0$, left plot), and a random walk from a starting constant ($\rho = 1$, right plot), splitting the series by countries with the highest or lowest values of ψ_i, and averaging over those values. Although countries with a high ψ_i always have more conflict than countries with a low ψ_i, in the left plot, both sets of countries have random year-to-year variation around their group mean. In the series with persistent shocks on the right, conflict appears to be following an upward trend. However, like the inverse-U trend in the instrument, the upward trend in conflict is a coincidence. Nothing about the data generating process ensures a positive trend, only smooth transitions that appear due to persistence create these

trends and cycles. The fact that this trend is stronger for some countries rather than others results from the scaling of the time series process underlying conflict by the country effect ψ_i.

Figure B2: Two draws of the dataset to contrast influence of persistence parameter on simulated outcome

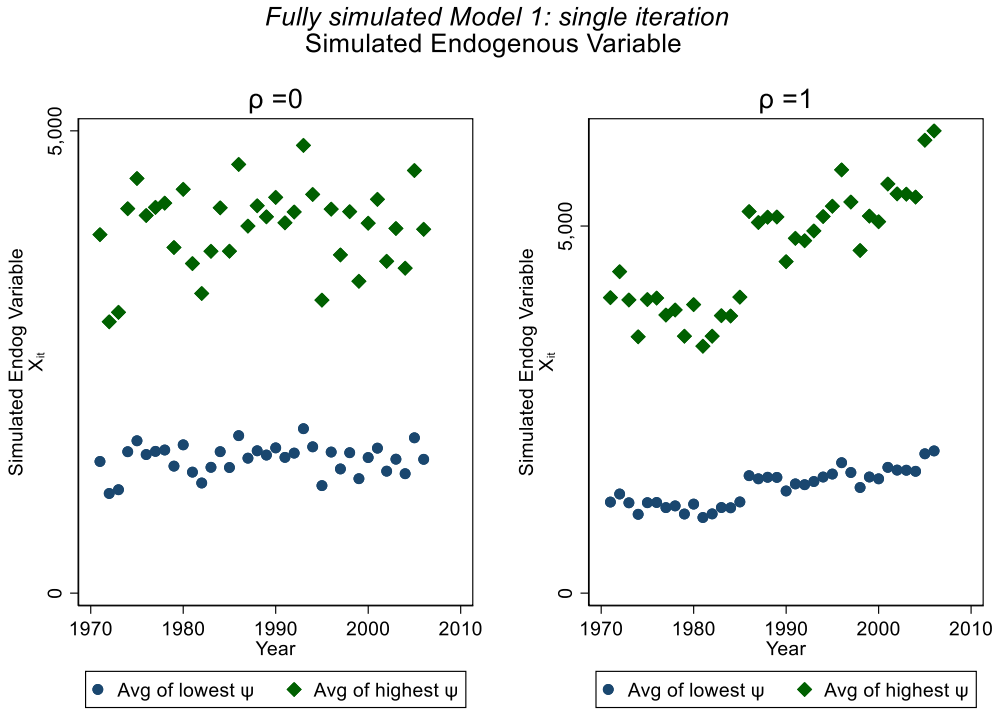


Notes: Data in these figures are from a fully simulated dataset with N=125, T=36 simulating random draws for equations B1-B9.

The plots below show the time series variation of the simulated endogenous variable X_{it}. Because the endogenous variable is strongly influenced by conflict, the trends appear similar to the conflict variable.

‘

Figure B3: Plotting two draws of the dataset to contrast influence of persistence parameter on simulated endogenous variable



Notes: Data in these figures are from a fully simulated dataset with N=125, T=36 simulating random draws for equations B1-B9.

Within a given simulated dataset like the one above we can see that we will generate trends and cycles in our variables of interest when the shocks to these variables persist over time. Within a single simulation however, a trend could be a coincidence, and we are interested in how these trending variables translate to IV estimates across simulations. In particular, we are interested in estimating coefficients from estimating five equations on the data generated by this system:

$$c_{it} = \gamma_0 + \gamma_1 Z_{it} + \epsilon_{it} \quad \text{(B10)}$$

$$X_{it} = \pi_0 + \pi_1 Z_{it} + \mu_{it} \quad \text{(B11)}$$

$$c_{it} = \alpha^{2sls} + \beta^{2sls} \widehat{X_{it}} + \eta_{it}^{2sls} \quad \text{(B12)}$$

$$c_{it} = \alpha^{OLS} + \beta^{OLS} X_{it} + \eta_{it}^{OLS} \quad \text{(B13)}$$

$$c_{it} = \alpha^{OLSFE} + \beta^{OLSFE} X_{it} + \theta_i + \delta * year_t + \eta_{it}^{OLSFE} \quad \text{(B14)}$$

In equation B10, the coefficient of interest is γ_1, which captures the reduced form relationship between conflict and the instrument. In equation B11, the parameter estimate of interest is π_1, the estimated first stage relationship between the instrument and the endogenous variable of interest. Equation B12 estimates β^{2SLS}, the IV relationship between X_{it} and c_{it} estimated by 2SLS by regressing c_{it} on $\widehat{X_{it}}$, the predicted value of $X_{it} = \widehat{\pi_o} + \widehat{\pi_1} Z_{it}$ from the first stage equation B11. To compare the IV relationships to the OLS ones, we also estimate B13 and B14, the OLS relationship between c_{it} and X_{it} with no controls (B13) and with country fixed effects and a linear year trend (B14). We then simulate this dataset 300 times and compare the estimated coefficients across simulations.

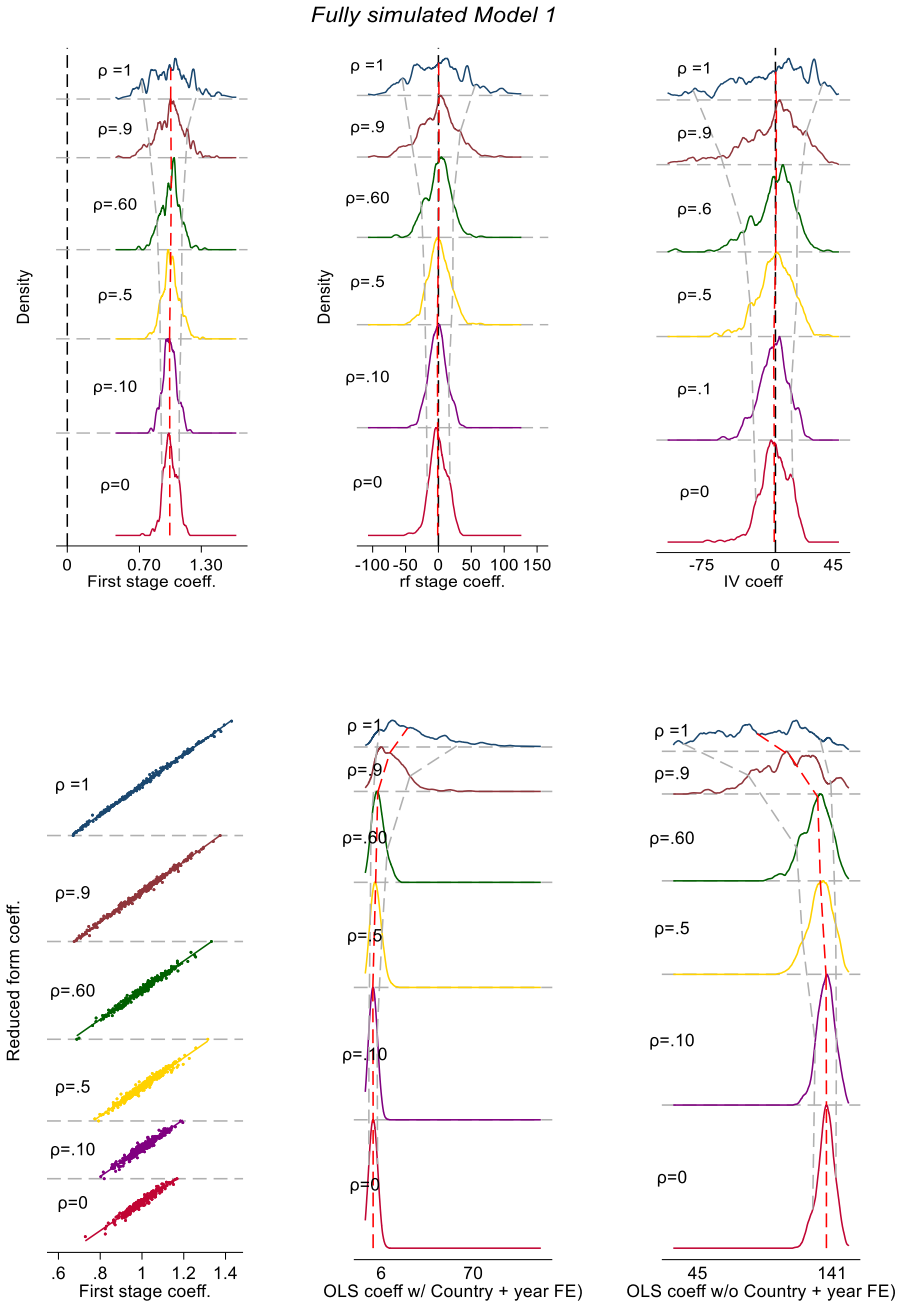
The left panel of the figure below shows the distribution of estimated first stage relationships between c_{it} and X_{it}, the estimated values of $\hat{\pi}$ from estimating B11 by OLS. The true first stage relationship is 1, as we know from equation B2 that an increase of one unit of the instrument increases c_{it} by 1. However, c_{it} also appears in equation B2 and could in principle mask the true first stage relationship. Because we set the coefficient on c_{it} near zero, the first stage coefficient estimates are (nearly) unbiased around one.

The persistence in the time series manifests in the first stage in the rising dispersion of the sampling distribution of the coefficient estimates across simulated samples as ρ approaches 1, as shown by the expanding 10th and 90th percentiles of realized coefficients depicted by dashed lines in the left panel. As in the familiar spurious regressions problem in time series, we become increasingly likely to estimate a very large positive or very large negative first stage as ρ increases.

The spurious regression problem also appears in the estimated reduced form equation, shown in the central panel. Across draws of the simulated datasets, the estimated $\hat{\gamma}$ is distributed around 0, meaning that median experiment correctly concludes that the there is no correlation between the instrument and the outcome in equation B1. However, as we increase persistence of the instrument and the outcome of interest, by increasing ρ from 0 to 1, we find an increasingly larger share of our simulated datasets return $\hat{\gamma}$ estimates that are farther from 0.

The right panel shows the implications for the 2SLS-IV estimate of β^{2sls} in equation B12. The estimated 2SLS-IV coefficients is approximately unbiased; they are centered on zero, meaning that in approximately half of our simulations, we find a negative relationship, and in half we find a positive one, which is reassuring when we know the true relationship is zero. However, the distribution of β^{2sls}estimates become more volatile as c_{it} and Z_t become more persistent. Thus when endogeneity is not a serious concern and the first stage is strong, we can still replicate the canonical time series spurious regressions problem in the panel IV estimator.

Figure B4: Distributions of estimated coefficients on simulated data



Notes: Data in this figure are distributions of estimated coefficients from equations B10-B14 estimated on a fully simulated dataset generated by the system of random variables described by B1-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets. Dashed lines show the 10th and 90th percentile of each distribution.

Per equation (7), the endogeneity of c_{it} and X_{it} in equation (B2) causes the finite sample correlation of τ_t and z_t to appear in both $\hat{\gamma}$ and $\hat{\pi}$, so that these coefficient estimates are highly correlated. This strong correlation of the first stage and reduced form is shown in the left plot of the bottom panel of Figure B4. Note that the shocks to τ_t and Z_t are uncorrelated by construction ($E[\text{cov}(s_t,q_t)] = 0$); Z_t and τ_t are independent processes. But the slow evolution of these variables when ρ is large cause the correlation between Z_t and τ_t to often be large in finite sample. This transmits the spurious regressions problem to the 2SLS-IV coefficient estimate. However, because the first stage is strong enough to always be positive (the x-axis is on the positive domain), the sign of the IV estimates is determined by whether the reduced form coefficient is positive or negative.

The OLS relationships in the center and right plots, always return a positive association between aid and conflict, whether or not controls for country fixed effects and a linear time trend are included. Together these results motivate the 2SLS-IV strategy. If the first stage is strong enough and the reverse causality weak enough, the 2SLS-IV estimates are much more likely to occur near the true null relationship than the OLS estimates, which are always biased and inconsistent. Not addressing the persistence of the main variables, however, will lead to mistaken inference. The appropriate remedy in this special case is to correct the standard errors using a Newey-West or similar heteroskedasticity-and-autocorrelation consistent (HAC) estimator that adjusts for autocorrelation in the error term.

B1.i. Does adding an interaction address the risk of spuriously large IV estimates in model 1?

A common strategy – one employed by both HI and NQ – is to control more flexibly for common trends by interacting the instrument with a variable that may be endogenous to the outcome in the cross section, but allows the inclusion of time fixed effects. To show how persistence affects this strategy, we create a new variable in the same simulated datasets used for Model 1 above, $\overline{X}_\iota = \left(\frac{1}{T}\right)\sum_{t=1}^{T} X_{it}$. When then estimate new first stage, reduced form and 2SLS specifications:

$$c_{it} = \gamma_0 + \gamma_1 Z_{it} * \overline{X}_\iota + \gamma_i + \gamma_t + \epsilon_{it} \quad \text{(B15)}$$

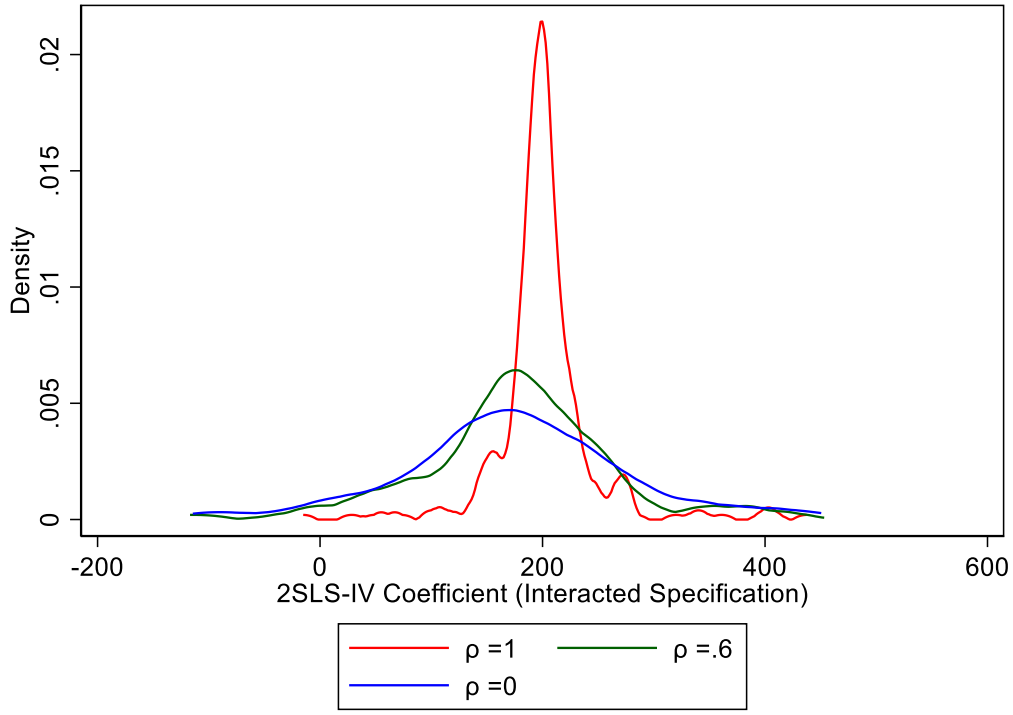
$$X_{it} = \pi_0 + \pi_1 Z_{it} * \overline{X_\iota} + \pi_i + \pi_t + \mu_{it} \quad \text{(B16)}$$

$$c_{it} = \alpha^{2sls} + \beta^{2sls-i} \widehat{X_\iota} * \overline{X_{\iota t}} + \alpha_i + \alpha_t + \eta_{it}^{2sls} \quad \text{(B17)}$$

The new cross-sectional variable $\overline{X_\iota}$ and the interaction of this variable in the first stage, reduced form, and second stage equations allows inclusion of both country and year fixed effects.[30] Using the parameterization of model 1, which minimized endogeneity bias, the figure below shows distribution of estimated $\widehat{\beta^{2sls-\iota}}$ from equation B17. Allowing for the inclusion of a nonparametric common trend through time fixed effects does not mitigate the program of volatility in the reduced form and second stage equations. Instead, this interaction strategy creates bias in this model, with nearly all realizations estimating a positive 2SLS-IV coefficient, and the problem where higher persistence leads to a more concentrated distribution around the incorrect, positive parameter value. Intuitively, this occurs because the IV coefficient estimate is a weighted average of the true causal effect and the bias arising from the spurious correlation of τ_t and Z_t. Since $\overline{X_\iota}$ is endogenous to ψ_i, the interacted terms increase the weight on $cov(\tau_t, Z_t)$, re-introducing the endogeneity problem through a different channel than we will show in model 2. This makes it all the more critical that any cross-sectional interaction term is itself fully exogenous. If a smoothly trending omitted variable affects conflict across countries and the variable of causal interest is endogenous to conflict, the interacted IV may exacerbate the inconsistency of the IV estimates by increasing the weight on the countries most affected by the trends in global conflict risk.

[30] This most closely resembles the NQ strategy of interacting US wheat production with the share of years in which recipient countries received US food aid. But it also informs HI's interaction strategy, where the interest rate time series is interacted with cross-sectional variables that may be strongly correlated with conflict, like ethnolinguistic fractionalization.

Figure B5: Distribution of 2SLS-IV coefficients from interacted specification estimated on simulated datasets



Notes: Data in this figure are distributions of estimated coefficients from equations B17 estimated on a fully simulated dataset generated by the system of random variables described by B1-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

B1.ii.Does taking first differences mitigate the spurious coefficient problem?

As a final check, we show that specifications that first differencing can help remedy the spurious regressions panel IV estimation problem in these simulations. For $\Delta c_{it} \equiv c_{it} - c_{it-1}$, we estimate the first stage, reduced form, and second stage equations in first differences as:

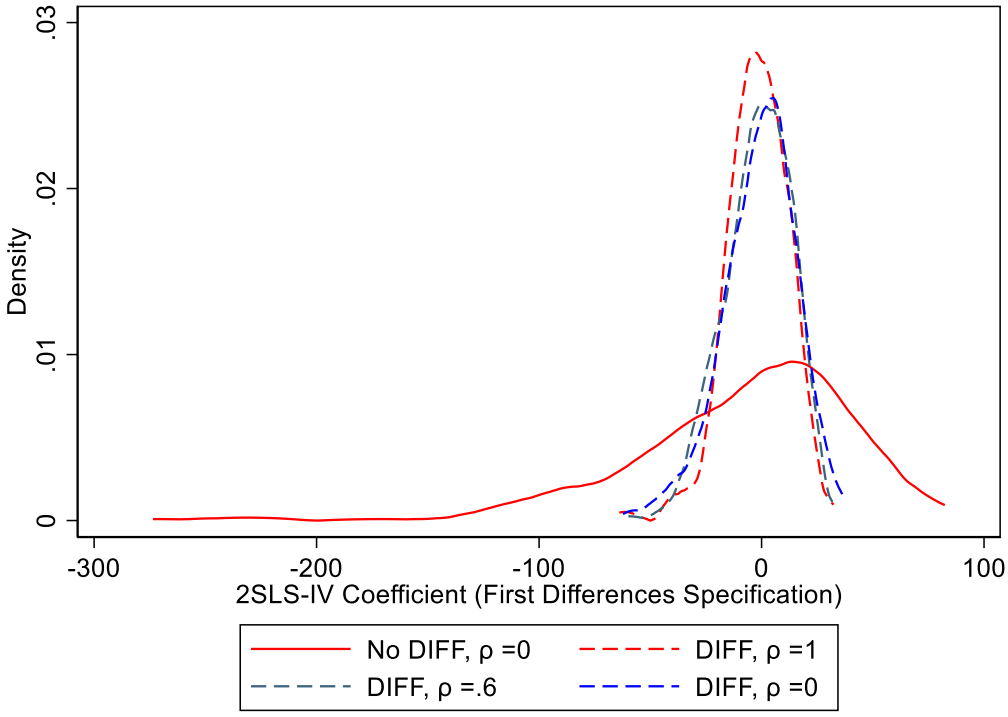
$$\Delta c_{it} = \gamma_0 + \gamma_1 \Delta Z_{it} + \epsilon_{it} \quad \text{(B18)}$$

$$\Delta X_{it} = \pi_0 + \pi_1 \Delta Z_{it} + \mu_{it} \quad \text{(B19)}$$

$$\Delta c_{it} = \alpha^{2sls} + \beta^{2sls-d} \Delta \widehat{X_{it}} + \eta_{it}^{2sls} \quad \text{(B20)}$$

The figure below shows the distribution of estimated $\widehat{\beta^{2sls-d}}$ when $\rho = 0$, $\rho = .6$, or $\rho = 1$, comparing against the distribution of estimate β^{2sls} without taking first differences on this same data. Unlike the interacted specification, the first differences specification does not reintroduce the bias arising from the persistent variables in the system, and both sets of coefficients are similarly distributed around the true zero value, without the excess mass of very large or very small coefficient estimates we get when $\rho = 1$ and we do not correct for persistence by differencing.

Figure B6:Distributions of 2SLS-IV coefficients estimated on fully simulated datasets with first differences specifications



Notes: Data in this figure are distributions of estimated coefficients from equations B20 estimated on a fully simulated dataset generated by the system of random variables described by B1-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

B1.iii. Would weak F statistics diagnose a spurious correlation problem in this model?

For all 300 simulations of this model, the F-statistic associated with equation the simple

2SLS specification (equation B12) or the first differenced SLS specification (equation B20) is always above 10 for every value of ρ. This is expected, because for this model, Z_t is a strong instrument in the sense that the correlation between X_t and Z_t is strong relative to other sources of variation in X_t. The problems from persistence when the first stage is strong and the source of concerning endogeneity is weak are related to the size of 2SLS-IV coefficients and are not diagnosed by weak instrument tests. In such a model, the value of corrections like differencing will primarily appear through reductions in the risk of estimating IV coefficients that are much larger or much larger than the true value.

B2. Model 2: Irrelevant instrument and strong endogeneity

The problem with relying on a HAC estimator is that one typically has no good basis for assuming that endogeneity is minimal. Indeed, if one assumes there exists negligible simultaneity bias, why use an IV estimator? To see the role that endogeneity plays when we fail to address persistence, we simulate this system by retaining the model from Model 1 but simply changing the parameterization of equation B2 to:

$$X_{it} = 0.5c_{it} + 0 * Z_t + \eta_{it} \quad \text{(B21)}$$

This model now imagines that we have a strong source of endogeneity, but reason to believe the instrument is not a valid one because it has no first stage relationship with the endogenous variable. Normally, we imagine we can avoid mistaken conclusions in this case, because it is always possible to observe and test first stage relationships. This model allows us to test whether persistent relationships can ever lead to us to falsely conclude on the basis of finite sample correlations that there is a non-zero first stage relationship, and if so, what conclusions would we make from the resulting 2SLS-IV coefficients.

Increasing persistence in the time series again increases the volatility of the coefficient estimates across simulated samples. As in the familiar spurious regressions problem, we become increasingly likely to estimate a very large positive or very large negative first stage as ρ approaches 1. This occurs because, as shown in the bottom left panel of the figure below, as the persistence of the main variables Z_t and τ_t increases, the

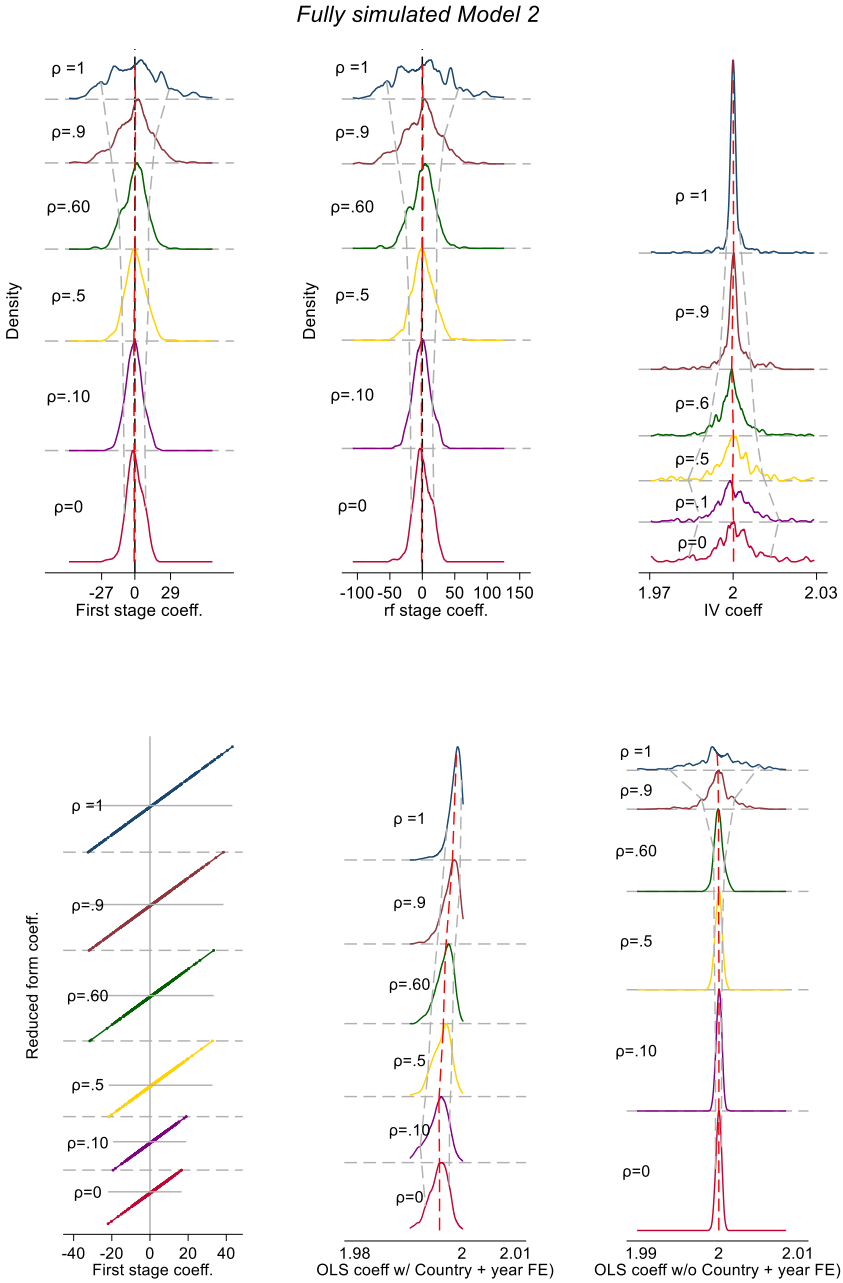
range of both first stage and reduced form coefficients expands, but the realizations across simulations always fall on the same upward sloping line. The consequence for the estimated first stage coefficients is that we are much more likely to estimate big first stage coefficients when variables are persistent over time. In this model where the true first stage relationship is zero, the 90th percentile of estimated coefficients across draws is over three times as big when $\rho = 1$ (28.7) as when $\rho = 0$ (9.1).

The spurious regression problem again appears in the estimated reduced form equation, shown in the top center panel of the figure below. Because the true relationship in equation B1 only contains one time series variable, persistence merely affects the standard errors, not bias. Across draws of the simulated datasets, the estimated $\hat{\gamma}$ remains distributed around zero, its true value. However, as we increase persistence of the instrument and the outcome of interest, by increasing ρ from 0 to 1, we find an increasingly larger share of our simulated datasets returning values of $\hat{\gamma}$ that are farther from 0.

The correlated spurious regressions in both the reduced form and first stage, however, generate bias in the 2SLS-IV estimates of β^{2sls} in equation B12. The fact that the reduced form coefficient and first stage coefficients are distributed around 0 does not mean that the 2SLS-IV coefficient will be distributed around zero as well. The strong correlation between $\hat{\gamma}$ and $\hat{\pi}$ estimates, shown in the lower left panel of the figure below, leads to increasingly concentrated sampling distribution around a (upwardly) biased estimate of β. From equation (14), we know that by assuming no causal effect of X_{it} on c_{it} in equation (B1) and a coefficient of 0.5 on c_{it} in equation (B2), we should find that the finite sample error introduced by the idiosyncratic errors s_t and q_t will generate $\widehat{\beta_{IV}} \sim \frac{1}{0.5} = 2$. Note that the sign and size of the bias in estimating the true beta will not depend on the degree of persistence in the time series. Indeed, as expected, we see in the upper right panel of the figure below that the estimated 2SLS-IV coefficients are centered around 2. In the presence of significant endogeneity, persistence again manifest in the variance of the $\widehat{\beta^{2sls}}$ sampling distribution. Persistence as simulated by a larger ρ causes the reduced form and first stage coefficients to each be more volatile. But because they are so strongly correlated,

the 2SLS-IV estimates become less volatile as persistence increases, in the sense that they are more tightly distributed around the biased estimate, in this case 2.

Figure B7: Distributions of parameters estimated on fully simulated datasets for Model 2



Notes: Data in this figure are distributions of estimated coefficients from equations B10-B14 estimated on a fully simulated dataset generated by the system of random variables described by B1-B9, replacing B21 for B2 (Model 2). Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets. Dashed grey lines connect the 10th and 90th percentiles of each distribution.

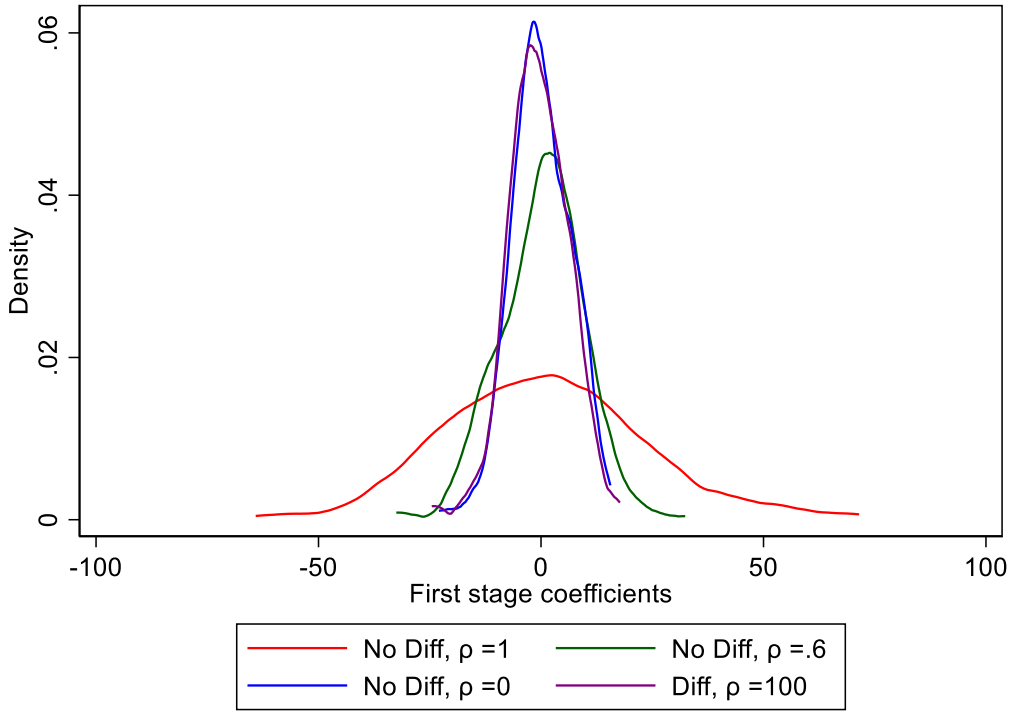
In practical terms, the model 2 simulation shows how endogeneity combined with time series persistence imperils panel IV estimation. This bias that arises from spurious regressions cannot be resolved with a HAC estimator, as that only addresses the standard errors, the mistaken inferences generated by spurious regressions.

Note as well that the reverse causality endogeneity the IV strategy aimed to address also implies that the instances when one falsely rejects a null first stage will also be the cases when one falsely rejects a null reduced form. The sign of the correlation of these estimates is also predictable in the sense that with a positive first stage, one also tends to get a positive reduced form relationship. As a consequence, the estimated 2SLS-IV coefficient is nearly always positive, no matter which direction we find a first stage or reduced form relationship. Incorrectly finding that a first stage is strong and significant on the basis of coincident time trends is not innocuous in the sense that the average of wrong experiments being right. Here all the bad experiments (irrelevant instruments) give the same wrong answer.

B2.i. Does adding an interaction or first differencing address the risk of spuriously large IV estimates in model 1?

If we estimate an interacted specification equations of interest (B16-B18) or take first differences (B19-B21), neither specification solves the finite sample bias in the estimated β^{2SLS}, for all values of ρ in both simulations, the 1st percentile of estimated $\beta^{\widehat{2SLS-\iota}}$ and $\beta^{\widehat{2SLS-d}}$ are always above 1.8, and the 99th percentile of both coefficients are below 2.2 for all ρ. However, when we always generate a (falsely) positive 2SLS-IV coefficient, the risk is falsely accepting the first stage relationship driven by the spurious time trends. The figure below shows that the distribution of estimated first stage coefficients from the differenced specification when persistence is high look like the coefficients from the non-differenced specification with low persistence.

Figure B8: Distributions of 2SLS-IV coefficients in fully simulated Model 2 datasets using first differences specifications



Notes: Data in this figure are distributions of estimated coefficients from equations B16 (No Diff) or B19 (Diff) estimated on a fully simulated dataset generated by the system of random variables described by B1, B21, B3-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

B2.ii. Do weak instrument tests diagnose the spurious correlation problem?

When $\rho = 100$, the KP weak instrument F statistic estimated for equations B10-B12 is greater than 10 in 88.0% of our iterations, and the p-value of the first stage coefficient is less than .05 in more than 99.3% of the simulations. This indicates that even when we know the true first stage is zero, spurious correlations appearing in the first stage that arise from having two time series processes in this equation would falsely lead us to conclude we had a valid first stage.

B3: Model 3: True non-zero first strange with stronger endogenous relationships than model 1.

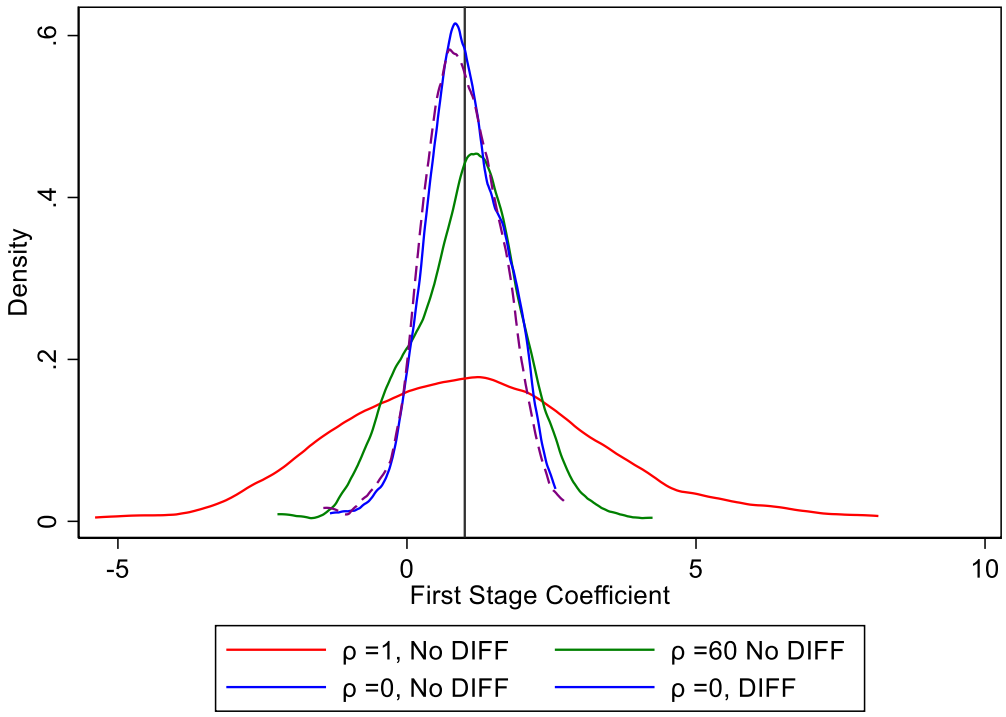
As a final model, we adjust the equation for the endogenous variable one more time:

$$X_{it} = 0.05c_{it} + 1 * Z_t + \eta_{it} \quad \text{(B22)}$$

As in model 1, there is now a true first stage in the data generating process as we have a positive coefficient on Z_t, but we have made the coefficient on c_{it} bigger than in Model 1. The value in this model is in combining the competing forces apparent in Model 1 (inflated 2SLS-IV coefficients when key variables are persistent even with a first stage) and Model 2 (endogeneity obscuring the true first stage when key variables are persistent). We choose parameters to show how persistence can cause us to estimate biased 2SLS-IV estimates with methods that would produce valid results in absence of persistence and in which stages differences appears.

The figure below shows the distribution of first stage coefficients ($\hat{\pi}$ from estimating equation B11). The true value of π is 1, and the coefficients across simulations are distributed around this true value, but persistence introduces volatility in the first stage. When $\rho = 1$, more than 30% of simulated datasets return a first stage coefficient with an incorrectly negative coefficient. Taking first differences reduces this volatility making the coefficients estimated with first differences (equation B20) approximately as closely distributed around the true value of 1with the persistent datasets ($\rho = 1$) as estimating equation B11 on a dataset with $\rho = 0$.

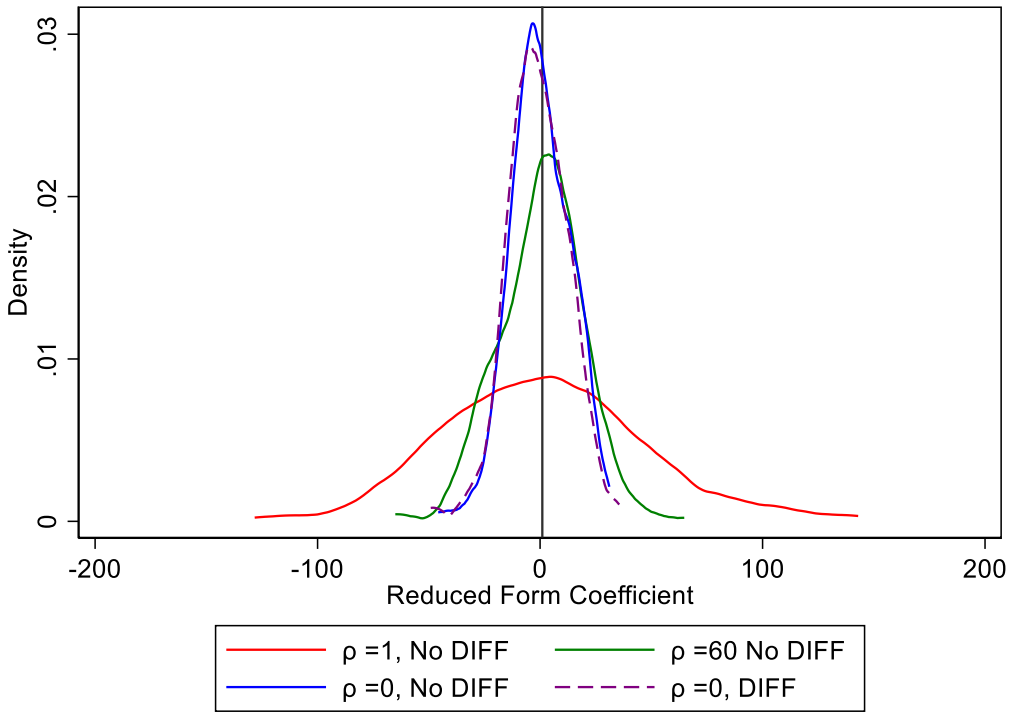
Figure B9: Distribution of first stage coefficients estimated on Model 3 simulated datasets with and without first differencing



Notes: Data in this figure are distributions of estimated coefficients from equations B16 (No Diff) or B19 (Diff) estimated on a fully simulated dataset generated by the system of random variables described by B1, B22, B3-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

The effect of persistence and the benefit of correcting this persistence in this case through first differences also appear in the reduced form coefficient. Coefficient estimates of $\hat{\gamma}$ from the reduced form (equation B10) are shown in the figure below. The distributions are centered on the true value of zero, but very large and very small values are more likely when estimated on a simulated dataset with high persistence ($\rho = 1$). As with the first stage, correcting for this form of persistence through first differences eliminates the effect of the persistence on volatility of the reduced form coefficient.

B10: Distribution of reduced form coefficients estimated on simulated datasets from model 3 with and without first differencing.



Notes: Data in this figure are distributions of estimated coefficients from equations B15 (No Diff) or B18 (Diff) estimated on a fully simulated dataset generated by the system of random variables described by B1, B22, B3-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

The risks of incorrect inferences emerge most clearly when plotting the 2SLS-IV coefficients for the simulated datasets resulting from this model. When $\rho = 1$ and the instrument and outcome both follow a random walk (red line left plot in the below figure), the 2SLS-IV coefficients are not distributed around the true value of 0. Instead, 16.3% of the simulations return a negative coefficient, and 84% return a positive one. When we take first differences (purple dashed line both plots below), the distributions are distributed around 0. The right plot shows that taking first differences returns a similar distribution of coefficients when $\rho = 0$ estimated by equation B12 or taking first differences as in equation B21.

B4: Summary of implications of persistence from models 1-3

Together these three models illuminate the role persistence plays in panel IV estimation. Even when drawn from independent processes, persistence in the instrument and the outcome makes the first stage and reduced form coefficients more volatile, in the sense that a dataset of the same size will be more likely to return a coefficient farther from the true value when ρ is relatively close to 1 than then ρ is relatively close to 0.

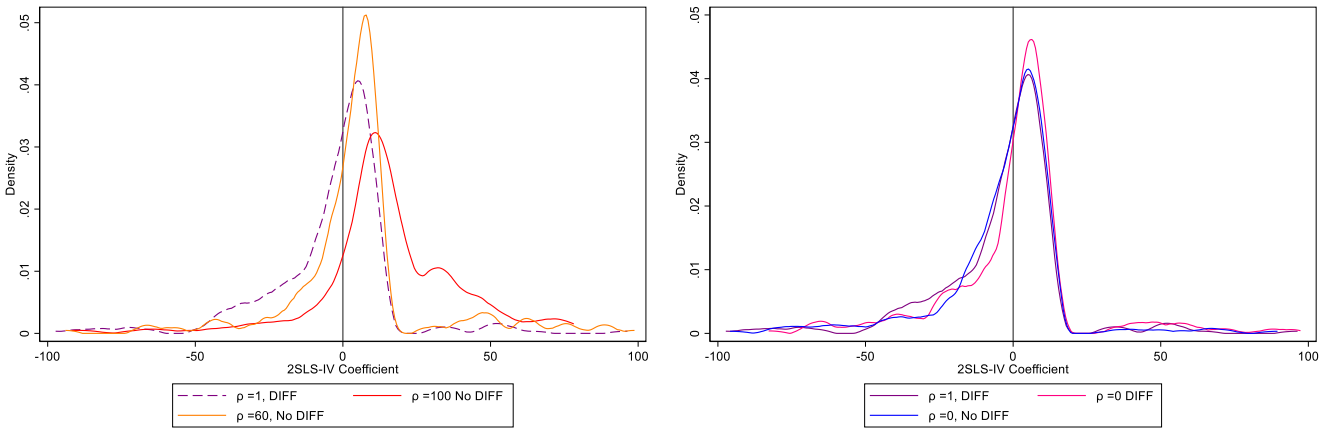
When the first stage is strong enough relative to the endogeneity as in Model 1, more volatility in the first stage from more persistence is not enough to change the sign of the first stage, and the share of IV coefficients that have the opposite sign of the true causal effect will be determined by the share of reduced form coefficients that have the opposite sign of the causal effect, which becomes more likely as we increase persistence and coefficients become more volatile.

When endogeneity is strong and the true first stage is weak or zero as in Model 2, volatility again arises in both the first stage and the reduced form as we increase persistence. In the extreme case of model 2, the volatility is distributed equally around 0 in both the reduced form and first stage. But the endogenous relationship between the first stage and the reduced form causes this random noise to always return a 2SLS-IV coefficient with the same sign as the underlying endogeneity. This problem is not solved by weak instruments checks. Many instruments will appear strong in the first stage because spurious correlations of the time series returns large, highly significant coefficients, but relying on these regressions does not guarantee a correct estimate of the causal effect, because the noise in the reduced form is systematically correlated with the noise in the first stage. Interacted specifications do not have the same benefit and always return positive 2SLS-IV coefficients for all levels of persistence on datasets simulated with this model.

Finally, we have shown through Model 3 the risk of generating coefficient estimates of the opposite sign of the true causal effect depends on the relative importance of persistence in the system. In cases where the variation in the endogenous variable

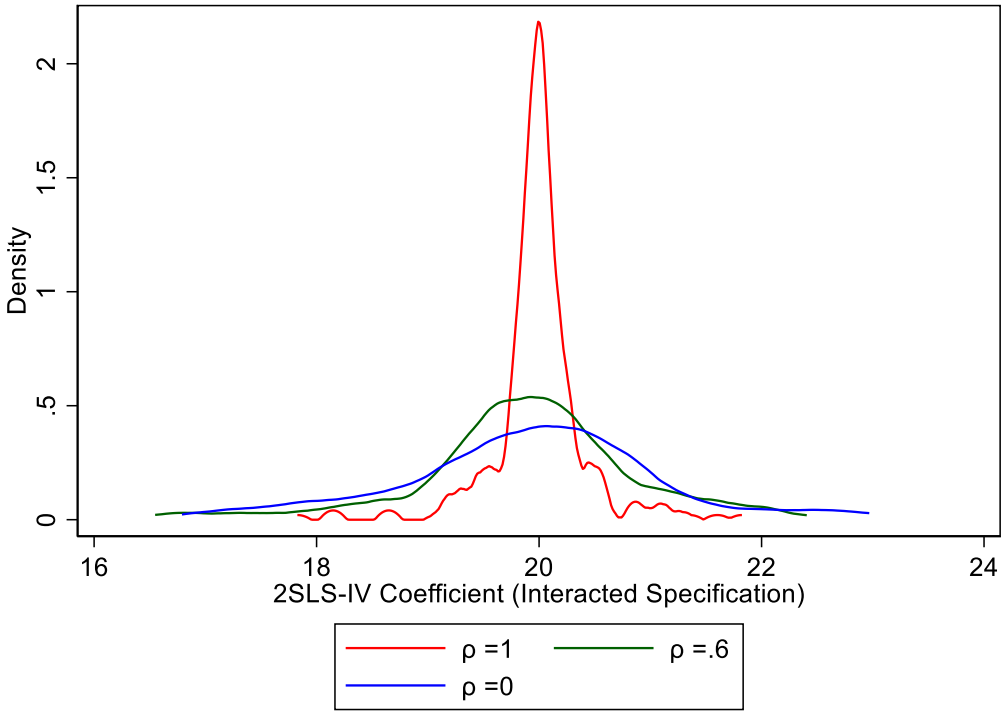
arising from the first stage or the endogenous relationship with the outcome are approximately balanced, the 2SLS-IV coefficients may be approximately correct when persistence is low, but systematically incorrect when the instrument and outcome variable approach a random walk. Corrections that remove this persistence, such as first differencing, reduce the volatility of both the first stage and reduced form and can therefore provide a helpful correction to the sampling distribution of the parameter estimate of interest.

Figure B11: Effect of first differencing on distribution of 2SLS-IV coefficients from model 3 for two levels of persistence.



Notes: Data in this figure are distributions of estimated coefficients from equations B17 (No Diff) or B20 (Diff) estimated on a fully simulated dataset generated by the system of random variables described by B1, B22, B3-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

Figure B12: Distribution of 2SLS coefficients estimated with interaction specification for simulated datasets in model 3.



Notes: Data in this figure are distributions of estimated coefficients from equations B17 estimated on a fully simulated dataset generated by the system of random variables described by B1, B22, B3-B9. Each plot shows coefficients from a regression estimated on one of 300 randomly generated datasets.

Appendix C: Simulation results with increasing time dimension

Given that finite sample correlation of unobservable determinants of conflict and the instrument cause many problems, we may want to know how quickly this term convergences to zero in typical sample sizes. The above Monte Carlo approach can be used to investigate consistency by re-running the same simulations with increasingly long time series. Table C1 reports the mean ILS IV coefficient estimated using 1,000 irrelevant random walk instruments when using shorter time series. The bottom row reports the bias when using the full conflict time series, all 36 years from 1971-2006. In each of the other rows, we start the conflict series in 1971, but end after 10, 20, or 30 years, respectively. Note that the pattern of bias that arises when using a shorter time series is irregular, not

even monotone in time series length. The simulations using the first 30 years of the data have a more biased distribution than the first 20, which has a more biased distribution than using only the first 10 years.

Table C1: IV coefficient using shorter time-series

Years	$\widehat{\gamma^{sim}}/\widehat{\pi^{sim}}$
1971-1981	0.0009274
1971-1991	0.0079909
1971-2001	0.0124657
1971-2006	0.0014036

Intuitively, this pattern arises because random walk variables often follow cycles, as Yule (1926) observed long ago. What matters, therefore, is not the duration of a time series so much as which portion(s) of the cycle one captures in the sample. Perhaps for the first ten years, the variables trend uniformly upward or downward. Therefore, including a linear trend as a control effectively absorbs this variation, eliminating most of the spurious correlation. But if the ten years instead captures a sub-period with a non-monotonic trend, the misspecification bias arising from correcting for a linear trend will increase rather than decrease as we add more years to the sample. The implication is that there is no substitute for inspecting the data for nonlinear trends. The bias does not disappear as one increases the number of periods within an intrinsically short time series.

Appendix D: Weak instruments tests in uninteracted models with simulated instruments and observed outcomes

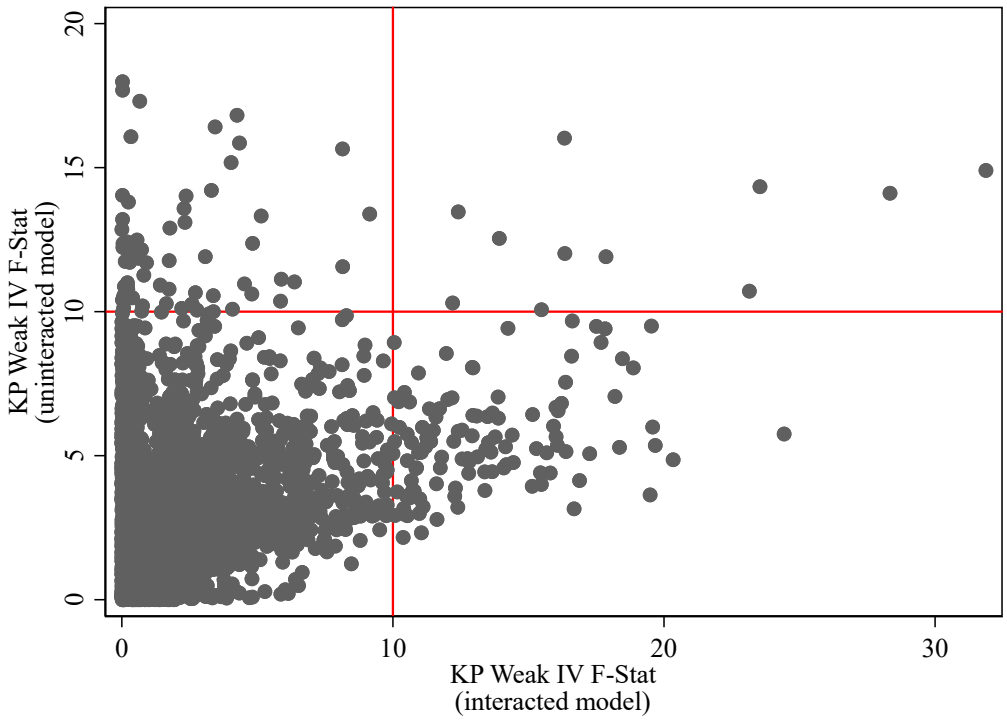
To understand the behavior of F-stats in simulations, we expanded the number of simulated instruments following a random walk to 3,500 in order to increase the observations of instruments with high F-stats.

In the uninteracted model, only 2.2% of simulated irrelevant instruments have an F-stat above the usual benchmark of 10, suggesting that this rule of thumb is a reasonable tool for distinguishing weak instruments from relevant ones. However, 29.8% of the F-stats for irrelevant instruments are above the value of 3.35 reported by NQ for their benchmark uninteracted specification.

Introducing a shift-share variable to the instrument generates problematic outcomes for the IV strategy. When the irrelevant variables are interacted with D_i, we now find that the pass rate for the weak instrument test threshold of 10 is 3.5%. This remains below the usual 5% threshold. But it signals that the interacted IV model increases the likelihood that weak instrument tests falsely conclude that irrelevant instruments are strong, in this case by 63% (from 2.2% to 3.5%).

This comparison understates the severity of the problem, however, because it does not account for the role of selecting the appropriate D_i variable. Authors can try any number of potentially endogenous D_i variables and check to see if the interacted instrument passes a weak instrument test with F>10. Once they find one that works, they can argue that the influence of that variable on conflict is absorbed by the country fixed effects, and that the interaction only adds power. The possibility of this sort of specification searching means that a key consideration is whether interactions simply make good instruments stronger, or whether strong instruments appear through the influence of the interaction rather than the plausibly exogenous time series instrument. In Figure D1, we show the scatter plot of F-statistics for the uninteracted and interacted weak instruments tests. Although F-statistics in the two models are correlated, they are only weakly so. This means that allowing the possibility of many different potential interactions increases the noise in weak instrument tests.

Figure D1: Correlation of Weak IV Test Statistics for Interacted and Uninteracted Models



Notes: F-stats are shown for the IV systems described by equations (26), (27) (Uninteracted) and (32), (33) (Interacted). Includes 3,500 simulated instruments merged to the NQ dataset.

To understand how the noisiness of this correlation affects the practice of implementing weak instrument tests, consider implementing one of several possible rules for whether or not to accept an IV as strong. First, suppose we accept instruments as valid only if they pass the weak IV test of F>10 in the uninteracted case. In our simulations, this would mean accepting as valid only 2.2% of the irrelevant instruments. Second, suppose we accept instruments as valid only if they pass the F>10 test in interacted cases. Then only 3.5% of the proposed irrelevant instruments would pass. Third, imagine that we accept instruments as strong if they pass the weak instrument test in *either* the interacted OR uninteracted model, as seems to be the current common practice. This rule would accept the instrument as valid in 5.4% of cases in our simulation. Allowing for interacted specifications thus reduces the power of weak instruments tests by half relative to considering only the interacted model. Expanding the set of possible interactions beyond

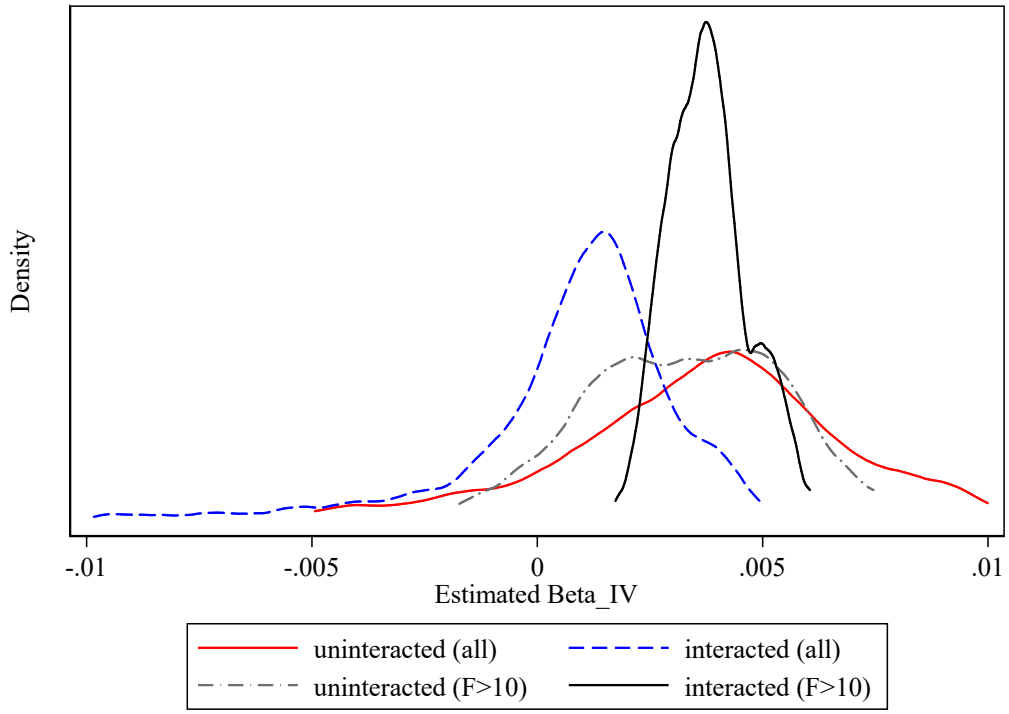
this one would reduce power further, because the weak IV tests will not be perfectly correlated across the different interactions.

Skepticism is warranted when a proposed instrument passes weak instrument tests in the interacted model and not the uninteracted model. Because the exclusion restriction is always justified by the time series variation of the interacted instrument, researchers are not typically expected to produce a theoretical justification for the excludability of the cross-sectional variable.

The other relevant consideration is whether using weak instrument tests help us avoid the bias that arises when we falsely accept an irrelevant instrument as valid. In Figure D2 we show the density of estimated IV coefficients when estimating the effect of aid on conflict using all 3,500 irrelevant instruments in an uninteracted model (red line), all 3,500 instruments in an interacted model (dashed blue line), only the instruments which return an F>10 in an uninteracted model (dashed grey line), and only the instruments which return an F>10 in an interacted model (dashed blue line). The comparison reveals that passing a weak instrument test does not avoid the bias arising from spurious time series correlations. Comparing coefficients estimated on irrelevant instruments that pass or do not pass tests (grey vs red lines), we see that similar bias emerges. But when comparing distribution of coefficients which pass or do not pass the weak instrument tests in the interacted models (blue vs black lines), we find that the distribution of IV coefficients among strong instruments only is more biased – and more concentrated around the incorrect parameter estimate – than the distribution of coefficients without strong instruments.

A final consideration is that weak or irrelevant instruments appearing strong is not the only concern caused by spurious correlations. As we showed in Model 1 of Appendix B, strong first stages can still generate misleading IV coefficients, because spurious correlation in the reduced form introduces error in the IV and can even reverse the sign of the IV coefficient if the spurious correlation is large enough.

Figure D2: Distribution of IV Coefficient Estimates Under Different Instruments



Notes: 2SLS-IV Coefficients are shown for the IV systems described by equations (26), (27) (Uninteracted) and (32), (33) (Interacted). Includes 3,500 simulated instruments merged to the NQ dataset.

The conclusion is that relying on weak instrument tests does not solve the spurious correlation problem because it reduces the power of weak instrument tests and increases the bias among the spurious instruments that do pass the weak instruments test.

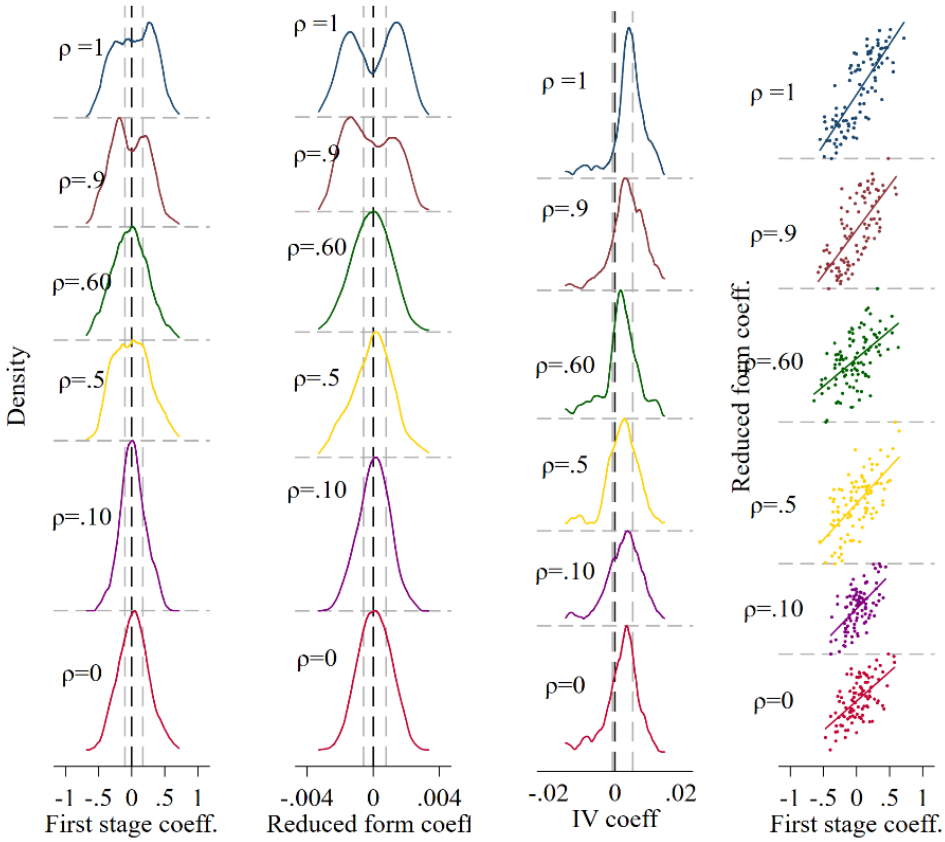
Appendix E: AR processes of different autocorrelation parameters in simulations with true outcome and exogenous variables

Figure E1 replicates the exercise in Figures 5 and 6, repeating each exercise with varying degrees of serial autocorrelation. The instrument in each simulation is generated by $Z_t = 100 + \rho * (Z_{(t-1)} - 100) + \epsilon_t$. The countries and years are held fixed by the sample used in NQ, and the outcome variable of interest is a dummy variable for any war in year t and country i as defined by NQ.

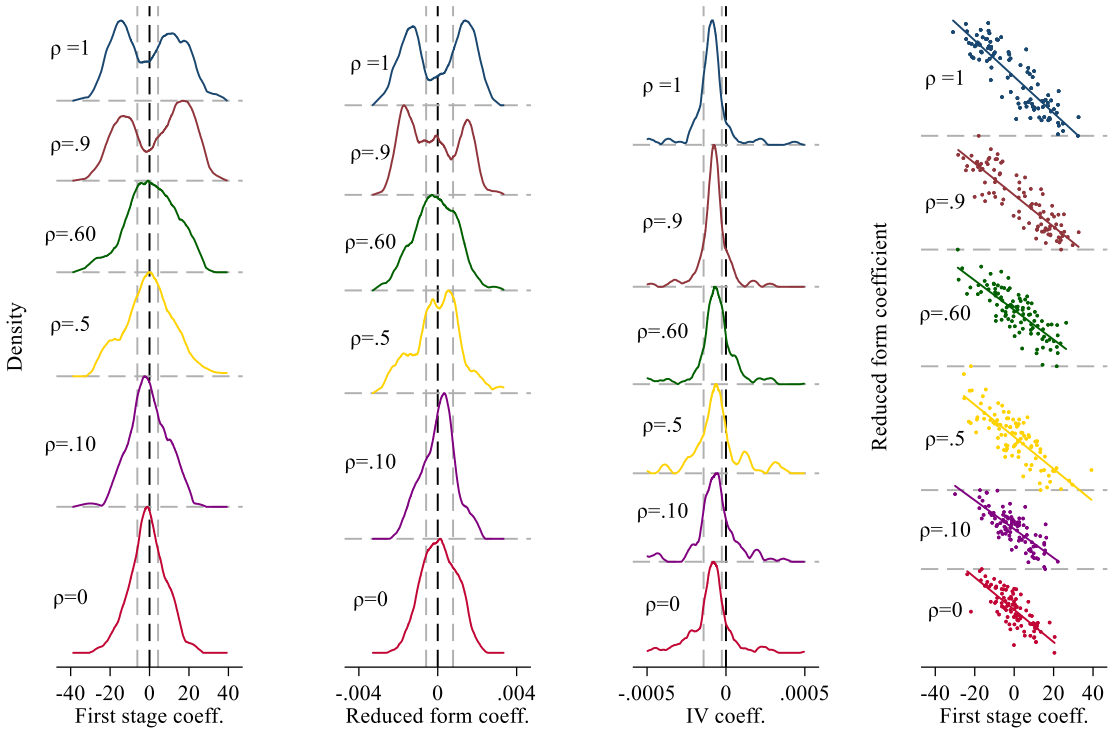
Comparisons of the distributions highlight the role of inference in the first and second stage and finite sample bias in the IV. Although the reduced form and first stage regressions are each separately unbiased as the expected value of coefficients across regressions is zero, the distribution of coefficients becomes diffuse away from zero as the degree of serial autocorrelation in the instrument increases. The IV estimate in the third column is always biased in the direction of expected endogeneity of the outcome variable conflict and the variable of interest (aid or conflict). The size of the bias is unaffected by the degree of serial autocorrelation.

Figure E1: IV estimates with simulated instruments of varying autocorrelation

a: Endogenous variable is wheat aid



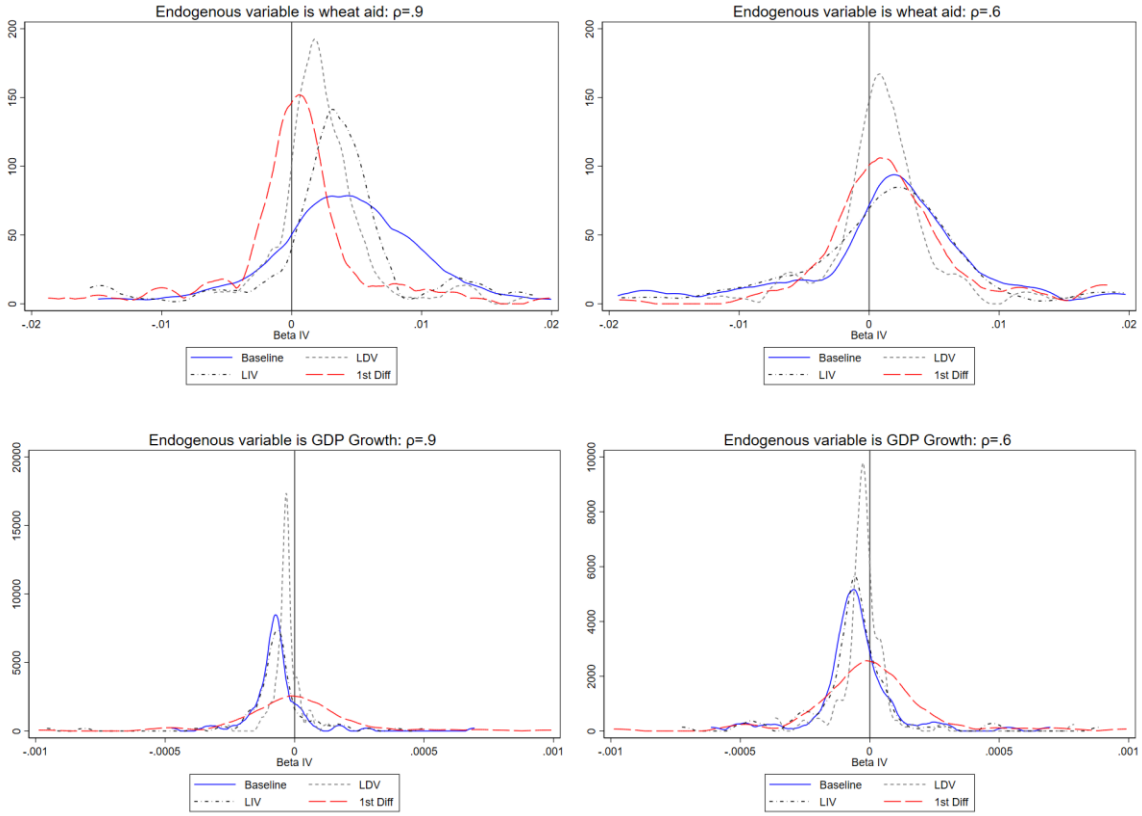
b: Endogenous variable is gdp growth



Notes: Each line is the density of coefficients estimated by 100 simulations. Densities are estimated by Epanechnikov kernel. The y-axis in the first three column is estimated density, in the fourth column, the y-axis is the reduced form coefficient.Instruments are fully simulated, outcomes, X variable (food aid), and controls are taken from NQ dataset and baseline specification. Dashed grey lines show the 25th and 75th percentiles of the distribution when $\rho = 0$.

To show how first differencing compares to other checks such as including a lagged independent or lagged dependent variable, we repeat Figure 7 for $\rho = .9$ and $\rho = .6$. The effect of first differencing on eliminating bias are most apparent for large values of ρ, but can still be seen in the centering of distributions around zero even when $\rho = .6$.

Figure E2: Comparing specifications for changing values of persistence



Notes: Each line is the density of coefficients estimated by 100 simulations. Densities are estimated by Epanechnikov kernel. The y-axis in the first three column is estimated density, in the fourth column, the y-axis is the reduced form coefficient. Instruments are fully simulated, outcomes, X variable (food aid), and controls are taken from NQ dataset and baseline specification.