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

By PAUL CHRISTIAN AND CHRISTOPHER B. BARRETT[*]

Abstract: Several recent empirical studies use instrument variables (IV) estimation strategies in panel data to try to identify statistically the causes of violent conflict. We explain how the long-recognized spurious regressions problem can lead to both bias and mistaken inference in panel IV studies given cycles in the time series component of the panel. We illustrate the problem by revisiting two recent, prominent studies that rely for identification on instruments exhibiting opposing cycles over time. Interacting endogenous cross-sectional variables with the time-varying instrument does not resolve the bias in general. When outcome variables are endogenous to interaction variables through policy preferences or reverse causation, the bias resulting from cointegration can be in the same direction as the reverse causation the IV is meant to resolve. We recommend practical diagnostic steps researchers can follow to reduce the prospect of spurious regressions confounding panel IV estimation. *Keywords: Instrumental Variables, Conflict, Foreign Aid, Economic Shocks, Panel Data*

[*] Christian: DECIE, World Bank (1818 H St, Washington DC 20433, email: pchristian@worldbank.org, phone: 2024738746, fax: 6072559984). Barrett: Charles H. Dyson School of Applied Economics and Management, 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, 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.

An important thread of quantitative social science strives to identify statistically the causes of violent conflict, a concern of first order importance.[1] As in so many other areas of empirical research, 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 5 of the studies Zürcher (2017) reviews 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 an instrumental variables (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 drivers of conflict beyond aid use similar panel IV methods in cross-country data to analyze non-aid prospective causes of conflict in multi-country data. A prominent example is Hull & Imai (2015, hereafter HI), who explore the impact of gross domestic product (GDP) growth on conflict. The estimation strategy HI and NQ (and others) employ relies on a plausibly exogenous time series instrumental variable to achieve causal identification.[2] In some specifications this variable is interacted with a (potentially endogenous) cross-sectional exposure variable(s) to allow controls for flexible trends through inclusion of year fixed effects.

In this paper, we show that seemingly-unobjectionable, exogenous time series instruments pose a problem, when one fails to expressly address the time series process

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

[2] 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 is meant as a caution and some practical guidance to those pursuing panel IV estimation, not as a critique of specific papers or authors.

underlying the instrument and the dependent and explanatory variables of interest. In particular, the NQ and HI papers and many others in the conflict literature – and the panel IV literature more generally – that use similar estimation strategies are vulnerable to the By ignoring the sequencing of observations in the panel and the resulting non-iid error processes of both the outcome variable and the instrument, inference that assumes iid errors leads to over-rejection of the zero-impact null. In this paper we go one step further to explain and demonstrate how in the panel IV context spurious regressions not only leads to mistaken inference, but also generates biased and inconsistent estimates. Indeed, in the special case of reverse causality between the outcome variable and an endogenous regressor, the resulting cointegration of these two variables introduces bias and inconsistency into the IV estimate that can reinforce rather than resolve the identification problem present in ordinary least squares (OLS) estimation and even generate estimates in the opposite direction of the true causal effect.

The problem we describe has both an inference dimension and a bias dimension. The inference side can perhaps be best understood by considering the likelihood that two random variables would be correlated by chance rather than through a causal connection, with reference to the time series properties of two series that are identical except for the temporal ordering of the observations. This simple exercise underscores the hazards associated with overlooking the temporal sequencing of observations in panel data estimation. Figure 1 below shows a variable Y plotted over a time series dimension t alongside a second variable Z.[3] In the left panel, both variables show substantial variation period to period, but in the panel on the right, they both follow smooth processes. If we simply regress Y on Z in both datasets, without reference to their time series properties, the correlation between Y and Z are the same in the two sets of data series shown in the two panels of Figure 1. In both series, Z tends to be high (low) when Y is also high (low). Using conventional inference methods, these correlations have the same t-statistics, confidence intervals, and p-values. However, the time series literature cautions

[3] To produce this figure with identical correlations, the series Y and Z take identical values in both simulations with Y following a known quadratic function over t. The difference between the two figures is only that the time series ordering of values in the righthand panel series is randomly shuffled in the left-hand series.

researchers about over-interpreting such correlations. Intuitively, the correlation in the data series in the left panel of the figure is less likely be caused by chance than is the correlation in the right panel. For inference purposes, if year to year shocks are like separate experiments, then a data series in which there are big shocks year to year simulates a series of well-powered experiments. Small shocks add little new information to the history of previous "experiments" generated by prior shocks. For identification purposes, it would be more difficult to think of an omitted variable that follows the same pattern as occurs in the left panel than in the right one. One naturally suspects that some omitted variable in the right panel leads to spurious correlation between Y and Z.

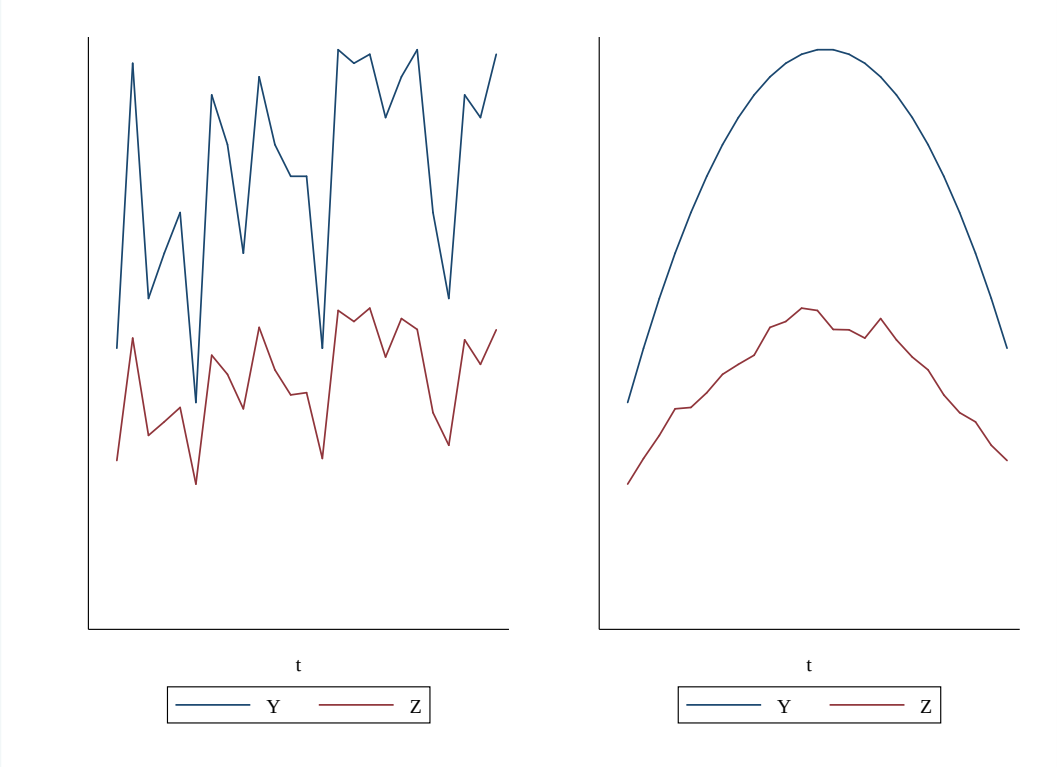
Failure to account for spurious correlations can play an important role in instrumental variables. Suppose Z and Y are generated from truly independent processes. and we are interested in understanding the effect of some variable X on Y. Then IV estimation using Z as an instrument for X can generate very different conclusions when the data series looks like the left panel of Figure 1 rather than the right panel. The inference problem inherent in a data structure like this is well known having been described in single time series by "spurious regression" critique Yule (1926), Slutzky (1937), and Granger and Newbold (1973) and in the panel data context by Bertrand et al (2004).

However, most IV applications – and arguments for the validity of the IV – treat these two cases shown by this thought experiment as equivalent. As we show, this has important consequences for the typical practice of asserting the validity of exclusion restrictions theoretically but testing first stage relationships statistically. The influence in an instrumental variables frameworks was shown by Phillips and Hansen (1990), who showed that in IV estimation involving cointegrated processes, the relevance condition of IV can sometimes be "satisfied" due to spurious correlation even when instruments are irrelevant due to independence from the endogenous regressor of interest.[4]When this problem is explicitly raised, it is often addressed through interacting with a cross

[4] The effect we study differs from the Nickell (1981) bias in dynamic panel data models that arises not from autocorrelated error processes but rather from the demeaning process, which generates inconsistency.

sectional variable. This practice relies on a common trends assumption that is highly suspect when the interacting variable is not also exogenous to the outcome. If Z is spuriously correlated with Y, it will also be correlated with any candidate instrumental variable X that is endogenous to Y, for example through reverse causation or policy preferences. This means that spurious correlation in the reduced form contaminates the first stage, interfering with our ability to assess the strength of an instrument's first stage.

Figure 1: Two equivalent correlated hypothetical data series



Note: The pairwise correlation coefficients between Y and Z, with no time series adjustments are 0.9847 in both panels.

In addition to flagging the ongoing relevance of an old literature on spurious regressions that has been largely overlooked in 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 using bootstrap and jackknife 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 problem arises because inference treats the dynamics of the instrument and other variables of interest as less smooth than they really are. Because this is a mirror image of problem studied by Young, we show that panel IV estimates that pass Young's test can still fall prey to spurious regressions. The problem we diagnosis is similar to other cases where inference that mishandles the smoothness of the distribution of errors leads to incorrect rejection. For example Kelly (2019) shows that inference is similarly mis-stated when authors fail to diagnose and account for spatial autocorrelation, which poses a similar problem to the spurious correlation in time series processes that we highlight.

In panel IV estimation, instruments can be constructed by interacting the time series variable with a cross-sectional shift or exposure variable. These interactions serve two purposes. First, in many cases, the interacted specification passes a weak instruments test while the uninteracted specification does not, meaning that the interaction terms allow the authors to report specifications that pass weak instrument tests. Second, the interacted specification can also bolster the argument that identification is not affected by misspecification of omitted trends because it allows for unobserved parallel trends. For example, in the two examples we study closely, 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 relatively exposed units to relatively unexposed units, in NQ's case, or mediating factors influencing the relevance of the instrument in HI's case. The logic is similar to the well-established shift-share instrument technique originated by Bartik (1991). 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.

Our paper thus complements recent critiques of estimation using interacted instruments, including the shift-share ('Bartik') IV strategy.[5] We clarify that identification assumptions arising either through exogeneity of shift-variables, as in Goldsmith-Pinkham et al (2018), or through exogeneity of shock variables, as in Borusyak et al. (2018), are onerous in the case of a single shock variable or highly correlated shocks. Our findings are perhaps closest to those of Jaeger et al. (2018), who point out a "dynamic adjustment bias" in panel IV estimates using shift-share instruments that arise from serially correlated processes in the literature on the wage effects of migration, and to Adão et al. (2019), who show via placebo simulations that conventional inference overstates rejection rates of null hypotheses in shift-share designs. Jaeger et al. (2018), studying the labor market effects of immigration, use a specific model of wage dynamics to illustrate a far more general point, that dynamic adjustments (i.e., convergence to new equilibria that takes longer than the time period associated with single observations in the data) can invalidate shift-share instruments by rendering them endogenous. We show that similar issues arise still more generally from spurious correlation that arises frequently due to time series properties that are common to many economic variables. Importantly, these concerns originate even without a specific omitted variable or adjustment process. As we show, the interacted instrument merely rescales the time series-based spurious regression bias that is our core concern. Thus, our critique extends well beyond the shift-share designs that have attracted much recent attention. This issue appears unrecognized in the literature, despite widespread use of these panel IV estimation methods, in the conflict literature and beyond.

In section I, we present a very simple model of heterogeneous trends to describe how the spurious time series problem can appear in instrumental variables estimation. This foreshadows the more specific concerns we demonstrate in greatly detail, and

[5] 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.

illustrate empirically, in subsequent sections. In section II, we describe the basic approaches employed to estimate the causal determinants of conflict using instrumental variables. Using the NQ and HI papers as examples, we describe the IV approach and show via regressions that replace the chosen instrumental variables with a clearly spurious variable how conventional inference tests can mislead. We then show through Monte Carlo simulations that autocorrelation in conflict leads to unreliable inference by traditional tests of statistical significance, and that IV estimates based on spurious instruments are biased. We investigate in our simulations the performance of robustness checks typically employed and show that correcting for the known time series properties of both the conflict variable and the instrument avoids problems of both inference and bias that arise when the relevance condition for IV is satisfied by spurious correlation. Using this insight, we re-estimate the NQ and HI specifications, showing that when afirst differences correction is used to correct for nonstationarity detected in the underlying data series, the results of the original papers are either overturned or become very imprecise, depending on the specification.[6] In section III we demonstrate that the use of interacted instruments in no way obviates the core problem, illustrating this result with the shift-share instrument case from NQ. In section IV, we propose seven practical diagnostic steps for how best to avoid spurious regression risk in panel IV estimation. These range from elementary steps – e.g., visually inspecting the data to identify the relevant variation in the time series or conducting well-established tests for nonstationarity – to more sophisticated diagnostics based on placebo tests and Monte Carlo simulation.. We close with some reflections on the state of empirical understanding of the causes of conflict and the formidable identification challenges researchers still face.

I. The Underappreciated Spurious Regressions Problem In Panel IV Estimation

[6] Chu et al. (2017) undertake robustness checks using a semiparametric endogenous estimation procedure and cannot reject NQ's parametric specifications, leading them to declare the original findings robust to alternative specifications.

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

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

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

The parameter of interest is the causal effect of X_{it} on $conflict_{it}$, β in this model. This simple model captures several key points challenges to estimating the causal determinants of conflict, and the conditions under which instrumental variables (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 both country-specific (ψ_i) and year-specific (τ_t) components may influence conflict, generating noise that might confound identification of β.[7] 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 may have a causal influence on X_{it}, but not on $conflict_{it}$. Initially, we consider an instrument that varies only in the time dimension, *t*, but we generalize this shortly.

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

$$conflict_{it} = \gamma Z_t + \mathrm{e_{it}} \quad (3)$$

$$X_{it} = \pi Z_t + u_{it} \quad (4)$$

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

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

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

[7] 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.

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 is simply the ratio of these two or:

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

As is well known, if and only if three conditions are all satisfied, then the ILS-IV estimator will yield a consistent estimate of β:

$$\operatorname*{plim}_{t\to\infty}\frac{\widehat{cov}(\tau_t,z_t)}{\widehat{var}(z_t)} = 0 \tag{8}$$

$$\chi \neq 0 \tag{9}$$

$$\widehat{var}(z_t) \neq 0 \tag{10}$$

Condition (8) is the standard exclusion restriction, requiring that the instrument only relates to the outcome variable through its influence on X.[8] Condition (9) is usually called the relevance condition; it requires that the instrument predicts variation in the X variable. As long as these conditions are met, and we observe variation in the instrument (condition 10), the ILS-IV estimator will converge to

$$\operatorname*{plim}_{t\to\infty}\frac{\hat{\gamma}}{\hat{\pi}} = \frac{\beta\chi}{\chi} = \beta \tag{11}$$

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.

This paper focuses on condition (8), the role of trends in the exclusion restriction. Note that because the instrument only varies along the time dimension, this condition turns on the correlation of two time series variables, the instrument Z_t and the time series component of the unobservable error, τ_t, in condition (8). The issues of 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. As we show, this has important implications for inference in that literature and familiar approaches do not routinely solve the problem.

[8] An alternative condition could replace condition (8), that $\bar{\psi} = 0$. This would require that year specific trends are symmetric across countries, so that they are on average zero. This specific case is rarely invoked and not relevant to the examples we consider.

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 $\operatorname{plim}_{t\to\infty} \frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} \neq 0$, and ILS-IV is not a consistent estimator for β. The fact that omitted trends can cause bias is fairly well known, even if the role of mis-specification is still often neglected, as evidenced by the fact that papers commonly report only one trend specification – typically linear – rather than a systematic approach to optimal trend specification.

Much less attention is given to the issue of spurious correlations arising 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. 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. This fact was demonstrated in simulations as early as Yule (1926), although proven analytically only recently by Ernst et al (2017). This problem, sometimes called the nonsense regression or spurious regression problem, has serious implications for estimating causal effects by instrumental variables in panel data. 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 predicable cycles. But when two truly independent variables each cycle, 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 have not 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 i.i.d.

To understand the consequence of this problem, consider how arguments for validity of an instrumental variable strategy typically proceed. When an instrument is proposed, the common rule of thumb is that the exclusion restriction is asserted but cannot be directly tested, while the relevance condition can and should be shown statistically through a strong first stage. A researcher typically argues that the instrument Z_t is externally determined and orthogonal to conflict. Then authors proceed by arguing that this exogenously determined instrument has an influence on X_{it} so it might satisfy the relevance condition. Even if it seems unlikely that X_{it} and Z_t are causally related, the argument goes, relevance can always be tested by estimating the first stage relationship and rejecting the null that $\pi = 0$ in equation (4). As long as appropriate care is taken with weak instruments tests, a finding that $\hat{\pi} \neq 0$, combined with a good argument for the conditional orthogonality of Z_t and $conflict_{it}$, usually suffices to establish the consistency of the ILS-IV estimate of β.

Time series correlations create a problem, however. 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 reduced form coefficient estimates (Equations 5 and 6). 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 $\beta = 0$, and imagine testing an instrument that is irrelevant in the true model ($\chi = 0$). If the instrument or the unobservable time trend follow a random walk or controls for trends are misspecified, $\frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)}$ will be large, and may not 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:

$$\hat{\gamma} = \bar{\psi} \frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} \tag{12}$$

$$\hat{\pi} = \alpha \bar{\psi} \frac{\widehat{cov}(\tau_t, z_t)}{\widehat{var}(z_t)} \tag{13}$$

$$\hat{\beta}_{IV} = \frac{\hat{\gamma}}{\hat{\pi}} = \frac{1}{\alpha} \neq 0 \tag{14}$$

The ILS-IV estimator is inconsistent for four reasons. First, if Z_t and $conflict_{it}$ are genuinely drawn from separate distributions, the exclusion restriction will seem plausible. But, second, the time series properties of these two variables will have a higher than

expected chance of generating a spurious correlation in any given sample, and standard statistical tests will overstate the significance of this association. Third, because this correlation enters the first stage through the simultaneous determination of $conflict_{it}$ and X_{it}, the first stage will *also* appear highly significant. This is perhaps the most overlooked pitfall of using time series variables as instruments. We cannot trust conventional tests of the relevance condition unless we have also checked and corrected all key variables in the first stage and reduced form estimates for time series issues such as trends and non-stationarity. Fourth, as seen in (14), even when the true effect of X_{it} on $conflict_{it}$ is 0, the ILS-IV estimate is $\frac{1}{\alpha}$. So spurious correlations arising due to common cycles (i.e., $\widehat{cov}(\tau_t, z_t) > 0$) will lead to upwardly biased IV estimates while counter-cyclical correlation (i.e., $\widehat{cov}(\tau_t, z_t) < 0$) will lead to downwardly biased estimates. When spurious time series correlation makes an irrelevant instrument appear relevant, the resulting distribution of IV estimates will not be centered around zero. In this simple model, they instead identify the inverse of the simultaneity coefficient from equation (2). The ILS-IV estimate is both biased and inconsistent in the same direction as the very source of bias that the IV was intended to solve.

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. A typical challenge is that trends of different functional forms can appear similar and tests to distinguish trends from a random walk are not well powered.

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

$$conflict_{it} = \beta X_{it} + \psi_i \tau_t \tag{15}$$

$$X_{it} = \alpha conflict_{it} + \chi_{int} w_i Z_t \tag{16}$$

Equation (16) follows the structure of a shift-share or Bartik instrument, interacting $w_i Z_t$. 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 demeaning all variables before

estimation. Estimating the first stage and reduced form coefficients on demeaned variables gives[9]:

$$\widehat{\gamma_{fe}} = \left(\frac{1}{1-\alpha\beta}\right) \bar{\psi} \frac{\widehat{cov}\left((\psi_i-\bar{\psi})(\tau_t-\bar{\tau}),(w_i-\bar{w})(z_t-\bar{Z}) \right)}{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))} + \frac{\beta\chi}{1-\alpha\beta} \frac{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))}{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))} \quad (17)$$

$$\widehat{\pi_{fe}} = \left(\frac{\alpha}{1-\alpha\beta}\right) \bar{\psi} \frac{\widehat{cov}\left((\psi_i-\bar{\psi})(\tau_t-\bar{\tau}),(w_i-\bar{w})(z_t-\bar{Z}) \right)}{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))} + \frac{\chi}{1-\alpha\beta} \frac{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))}{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))} \quad (18)$$

$$\hat{\beta}_{feIV} = \frac{\widehat{\gamma_{fe}}}{\widehat{\pi_{fe}}} = \beta + \frac{1}{\alpha} \frac{\widehat{cov}\left((\psi_i-\bar{\psi}),(w_i-\bar{w})\right) cov\left((\tau_t-\bar{\tau}),(z_t-\bar{Z})\right)}{\widehat{var}((w_i-\bar{w})(z_t-\bar{Z}))} \quad (19)$$

Fixed effects estimation creates new, less restrictive conditions to achieve consistency in estimating β. If any of the following conditions are true, then the time series correlation between $cov((\tau_t - \bar{\tau}), (z_t - \bar{Z})/ \widehat{var}((w_i - \bar{w})(z_t - \bar{Z}))$ will not enter the estimates of the first stage, reduced form, and thus the fixed effects ILS-IV estimate $\hat{\beta}_{feIV}$ will be consistent:

$$\psi_i = \bar{\psi} \ \forall \ i \quad (20)$$

$$\tau_i = \bar{\tau} \ \forall \ i \quad (21)$$

$$\widehat{cov}((\psi_i - \bar{\psi}), (w_i - \bar{w})) = 0 \quad (22)$$

Condition (20) is analogous to the common trends assumption in differences-in-differences estimation. Unobservable trends can affect conflict, but as long as these trends affect all countries equally, the inclusion of year fixed effects removes the influence of these trends. This condition is related to the exogeneity of shocks assumption described by Borusyak et al (2018). The second assumption rules out the time series problem by assuming that unobservable influences of conflict vary only across countries and not over time. The third condition (22) is weaker still. It allows for the possibility that unobserved trends are heterogeneous, but requires that this heterogeneity be uncorrelated with the cross-sectional variation in the instrument. This assumption is akin to the exogeneity of shares condition described by Goldsmith-Pinkham et al. (2018). Unfortunately, these assumptions do not appear widely understood in empirical applications.

The result of these conditions is that when we allow for the possibility that different countries or different sub-samples of countries follow different trends, the consistency and

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

unbiasedness of the IV estimator depends on the heterogeneity in trends being uncorrelated with the cross-sectional component of the instrument. Such correlations can easily arise, however, for example when the variable interacted with the instrument is endogenous the outcome variable (conflict).

When $(\psi_i - \bar{\psi})$ and $(w_i - \bar{w})$ are correlated, then the fixed effects estimator will not only preserve the spurious correlation arising from the time series variables, it also changes the estimated variance of this term. This creates a new threat to establishing causal identification: specification searching for apparently strong instruments. Researchers can propose time series instruments that arguably satisfy the exclusion restriction. They can propose cross-sectional variables to interact with the time series variable and find one that passes a weak instrument test. They then argue that their preferred instrumental variable is excludable because they control for country fixed effects. But per condition (21), level effects are not the identifying assumption to remove the influence of spurious trends; zero covariance is the key assumption. Researchers must argue either that there is no heterogeneity across countries in trends, or that the heterogeneity is not correlated with the cross-sectional interaction variable. Although similar concerns are familiar in differences-in-differences estimation, we are unaware of empirical papers that make this case in the context of interacted instrumental variables in panel data. As we show below, failure to satisfy this stronger condition leads to inconsistent and biased parameter estimates.

In the following sections of the paper, we use prominent papers from the conflict literature to show how trends appear in a commonly used conflict dataset. Using a clearly spurious variable and swapping instruments across cases, we show that the data exhibit trends that are correlated with the cross-sectional component of the interacted instruments. This precludes assessment of which variables are and are not valid instruments using common specifications and rules of thumb for instrumental variables. To illustrate the problem in a controlled setting, we then run a placebo test where we generate a variable from a random walk and employ it as an instrument for conflict. We show that this variable, known to be irrelevant, generates results remarkably similar to those reported in the published cases we consider, exactly as one would expect in the presence of spurious

trends. We next show that trends in the conflict data are heterogeneous and the dimensions of heterogeneity are correlated with the cross-sectional variation used by authors to create interacted instruments. Finally, we recommend a set of diagnostics that can be used to assess whether trends and non-stationarity are likely to confound panel IV estimation.

II. Panel IV Methods Without Interactions in Applications

We begin by ignoring for the moment 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 relationship of the following type in its core bilateral relationship:

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

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.[10] 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 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.

[10] USAID states explicitly: "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).

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.[11] 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. 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. 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:

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

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

Equation 3 is first stage, estimating the causal effect of the exogenous instrument, Z_t, on the endogenous regressor, X_{it}. Equation 2 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 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.

a. 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

[11] HI follow Shambaugh (2004) in classifying countries whose currencies are not explicitly pegged to another country's currency via a fixed exchange rate.

real interest rates) – and a rich set of characteristics of countries and years that the original authors use as controls.[12] When one looks at simple scatter plots of data, ignoring the temporal sequencing of observations, the panel IV identification strategy seems to work. Figure 2a shows the correlation between real interest rates and conflict, the reduced form relationship in HI. Interest rates and conflict covary positively. Figure 2b 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.

Figure 2a: Conflict and real interest rates

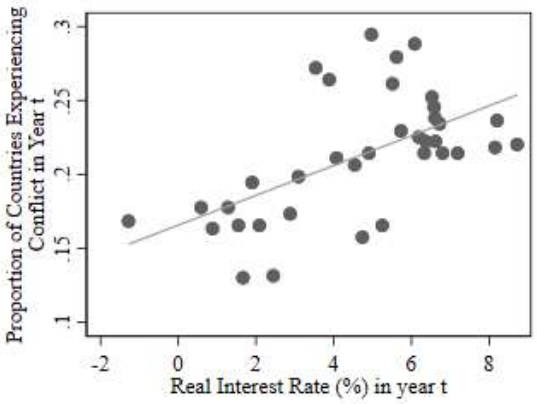


Figure 2b: GDP growth and real int. rates

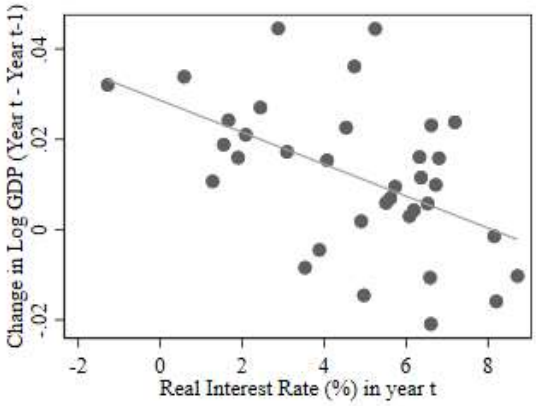
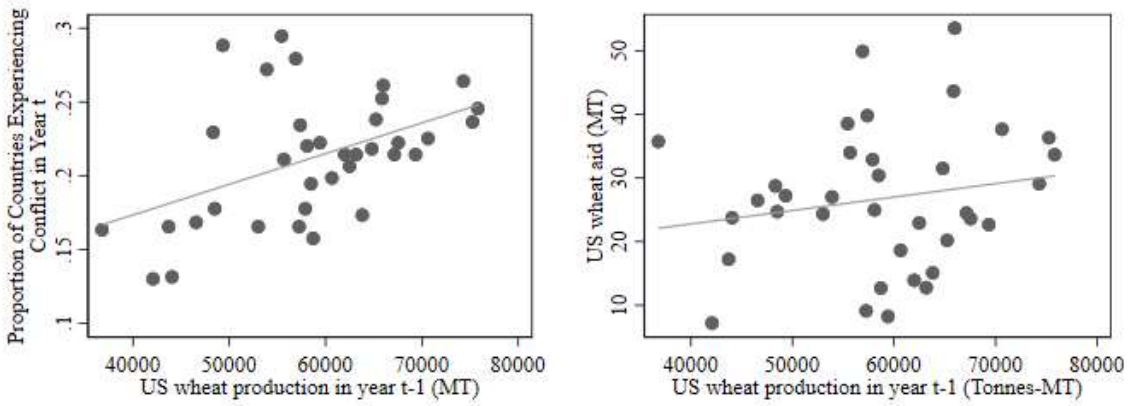


Figure 3a:
Conflict and lagged US wheat production

Figure 3b:
Food aid and lagged US wheat production

[12] 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.



Similarly, Figure 3a shows the positive reduced form relationship in NQ, between conflict and lagged US wheat production, while Figure 3b 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.

b. 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 actual data come from a specific time series. One could scramble the time series observations without changing the plots in Figures 2 and 3 and parameter estimates based on the relationships depicted in them at all.

Figure 4 displays the actual trends in the time series, estimated nonparametrically by lowess, in conflict (upper left panel), wheat production (upper right panel), interest rates (lower left panel) and a fourth variable, global audio cassette tape sales (lower right panel). [13] We chose the audio cassettes variable specifically because it is obviously spurious but exhibits a clear, nonlinear trend.[14] No coherent, credible mechanism exists that causally links audio cassette tape sales to conflict, real interest rates, or US food aid shipments.

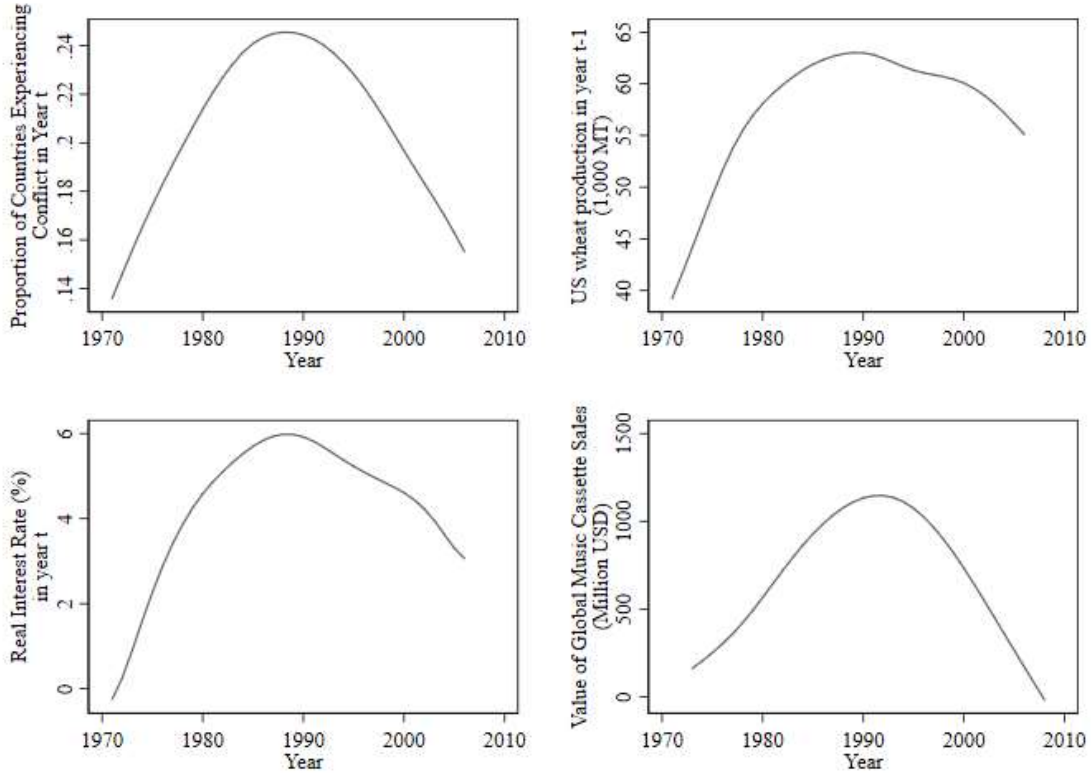
[13] The global audio cassette sales data come from IFPI (2009).

[14] 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.

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. 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 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 4: Underlying trends in the conflict and instrumental variables



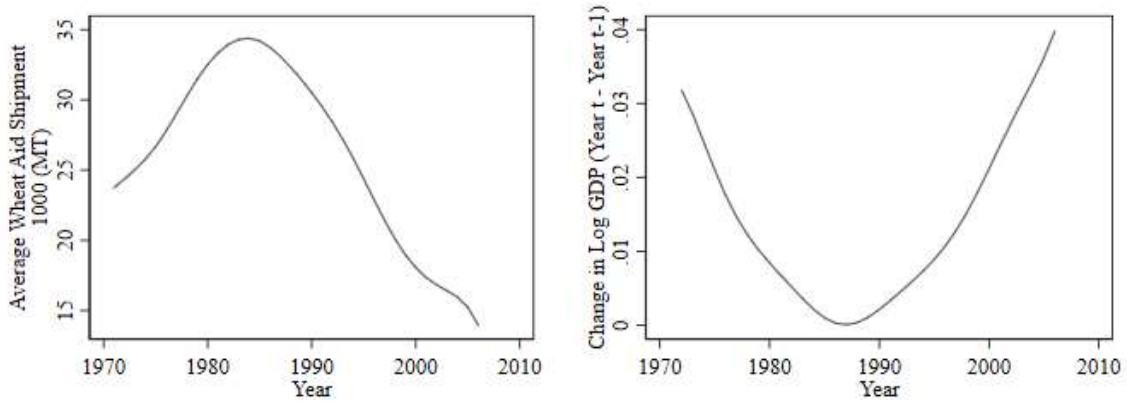
The relationship we care about is not the reduced form, but rather the relationship between the outcome (conflict) and the potentially endogenous explanatory variable (shipments of food aid or GDP growth). A reduced form relationship between an outcome and an instrument is only one criteria to check 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 2 and 3 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 5 displays the trends in the endogenous regressors of interest: wheat aid in panel 5a and GDP growth in panel 5b. 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 4, 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.

Figure 5a: US wheat food aid shipments *Figure 5b: GDP growth rates*



Given that our candidate instruments all have inverted-U trends, Figures 4 and 5 tell us what we already knew from Figures 2 and 3, that interest rates will be negatively correlated with GDP growth and that lagged US wheat production will be positively correlated 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.[15] A common trend among the dependent, endogenous explanatory, and instrumental variables means

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

that spurious and truly causal relationships will exhibit identical patterns, calling into question the causal identification.

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 estimates that are always highly statistically significant. Indeed, for the food aid regressor NQ study, the most precisely 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? – accounts for the observed correlations.[16]

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

IV =	**Dependent variable = incidence of war (of any type)**					
Endogenous regressor	**(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

First stage inference tests do not help us identify the spurious correlation. When 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 are 32.4 and 10.1, well above the standard threshold values of 10.

[16] 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.

The existence of multiple candidate instruments, at least some of which are almost surely spurious, in no way negates the possibility of a truly causal relationship between the endogenous regressor and a suitable instrument. But the process serves as a caution highlighting how apparently strong associations can arise through mishandling of IID assumptions in standard OLS and IV. The risk of erroneously accepting a spurious correlation is especially high because, as we show in the next section, traditional inference will over-reject the no-impact null in the presence of co-trending variables and will almost surely lead to biased 2SLS estimates in panel IV regressions. As we will go on to explain, the possibility of spuriously correlated trends makes it essential that researchers have in mind a specific, credible mechanism that relates the instrument they choose to the endogenous regressor, and that they subject that hypothesized mechanism to explicit testing. In section III we illustrate how one might do that, with reference to the celebrated NQ paper, which articulates a specific mechanism relating lagged US wheat production to US food aid shipments.

c. How correlated cycles affect panel IV inference: A Monte Carlo analysis

The previous section demonstrates the possibility of spurious correlation in the time series, but that possibility in no way establishes that the HI or NQ results are wrong. 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, which we now explore in more depth.

In this sub-section we use Monte Carlo simulation to investigate how autocorrelation in time series variables can cause both mistaken inference and bias. 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 build directly on those insights.

We begin by simulating an instrument that follows a random walk process.[17] Specifically, in each round we implement the following procedure, 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:

1. Define an instrumental variable Z_t that takes a value of 100 in year -36.[18]
2. In each subsequent year, there is a random shock that is uniformly distributed, $q_t \sim U(-0.5,0.5)$. 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.
3. In years 1-36, holding conflict, food aid flows, and all of NQ's controls from their baseline specification constant, we estimate the first stage, reduced form, and 2SLS equations from NQ, 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.
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 6a plots the distribution of the π^{sim} coefficients estimated in each of 1000 replications of the following first stage regression (suppressing included controls):

$$Aid_{it} = \pi^{sim} Z_t + \eta_{it}^{sim} \quad (4)$$

[17] The source of the problem is not specifically I(1) processes. One could also conduct this exercise with a variable that follows a trend stationary process instead. But random walk, I(1) processes are common in economic variables, intuitive, and allow us to show that these problems need not arise from any specific omitted variable or deterministic trend. The problem arises simply from the smooth dynamics of the instrument when some information from past realizations is preserved in contemporary ones.

[18] We use 36 periods simply to replicate the duration of the sample used in estimation. This is inherently arbitrary. We just wanted to start with a deterministic expected value and have generated random sequences of the time series value from a known expected value.

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).[19] 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.

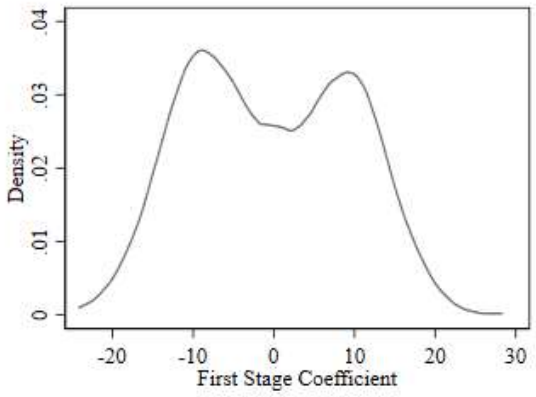
This bimodal distribution of the reduced form coefficients implies that large reduced form relationships between conflict and irrelevant instruments will arise more often than conventional inference tests would suggest. This is the practical implication of Yule's "nonsense correlation" problem, later often referred to as "volatile" or "spurious" correlations. It arises because, as shown by Slutzky (1937), variables accumulate a past history of random shocks, and the sum of random shocks can always be modeled as a cyclical sequence.[20] A robust time series econometrics literature, especially since Granger and Newbold (1974), recognizes the spurious regression problems that arise when regressing one variable that follows a random walk on another that also follows a random walk. Hence the standard time series practice of testing for difference and trend stationarity in variables on both sides of a regression before estimating correlations between them. In the next section, we apply these conventional time series tests to the NQ and HI variables. But first we show what this problem means in an IV setup with both a first stage and second stage equation rather than simply a single regression equation. We show that in the IV estimation setting, the problem moves beyond mistaken inference due to over-rejection of the null hypothesis to one of bias and inconsistency.

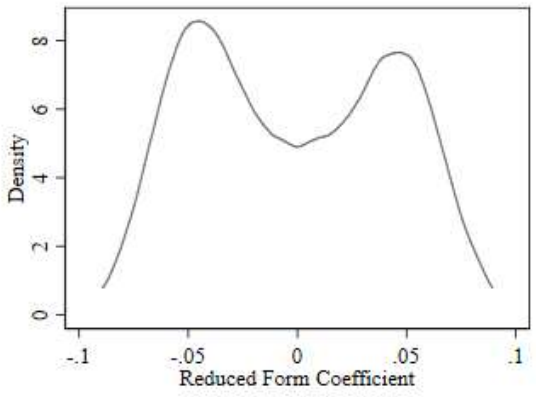
[19] Yule's (1926) empirical finding remains a subject of analytical research in statistics. Ernst et al. (2017) recently proved the result that the distribution of estimated correlation coefficients between two independent time series will be heavily dispersed. There do not yet appear to be analytical 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.

[20] Or as Slutzky (1937, p.105) more eloquently put it "Almost all of the phenomena of economic life, like many other processes, social, meteorological, and others, occur in sequences of rising and falling movements, like waves. Just as waves following each other on the sea do not repeat each other perfectly, so economic cycles never repeat earlier ones exactly either in duration or in amplitude. Nevertheless in both cases, it is almost always possible to detect, even in the multitude of individual peculiarities of the phenomena, marks of certain approximate uniformities and regularities."

Figure 6: Monte Carlo panel IV estimates with positively 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*

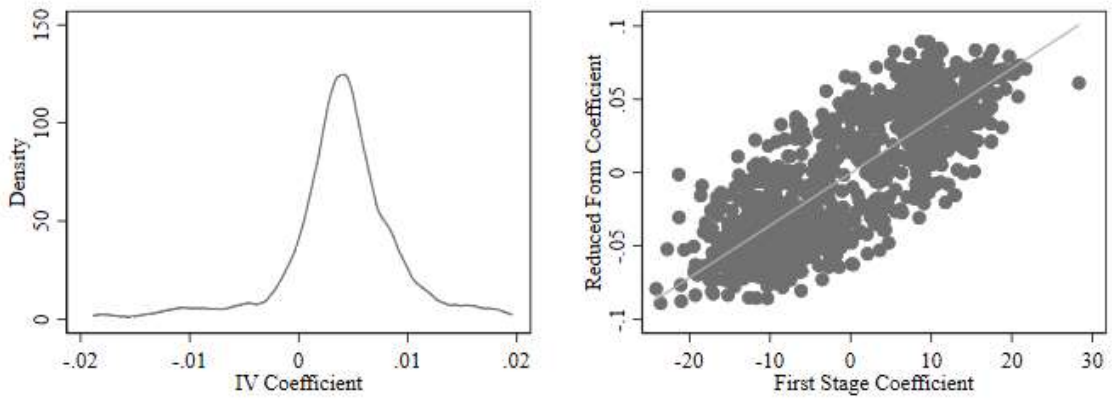


Figure 6b shows the distribution of coefficient estimates from the reduced form equation (again suppressing controls included in the simulations):

$$Conflict_{\text{it}} = \gamma^{sim} Z_{it} + \mu_{it}^{sim} \quad (5)$$

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 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 problem in either the first stage or the reduced form regressions is one of inference, not bias. The estimated $\widehat{\pi^{sim}}$'s in Figure 6a and $\widehat{\gamma^{sim}}$'s in Figure 6b center around zero, confirming the unbiasedness and consistency of the parameter estimate. With a sufficient number of 36-year samples, E[$\widehat{\pi^{sim}}$] =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 collect enough data, we will eventually get

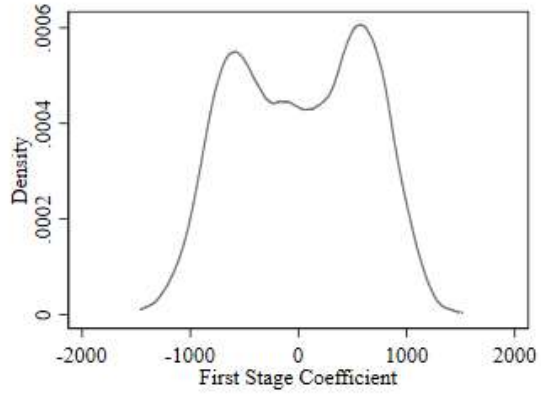
the right answer, even if the convergence might take a bit longer than conventional tests suggest.[21]

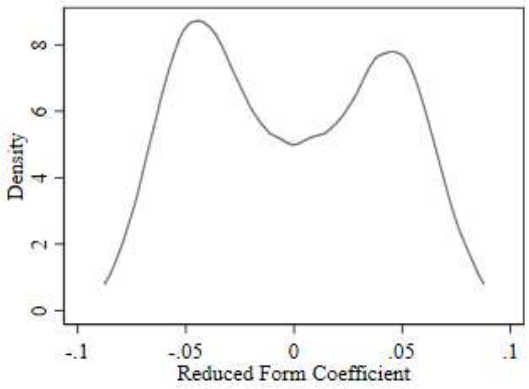
The greater concern is that the consistency and unbiasedness that holds for the OLS estimate estimated in the first stage or in the reduced form equation do not hold for the 2SLS/ILS estimate. The empirical distribution of the 2SLS estimate, shown in Figure 6c, is clearly positively biased and not centered around zero, which is implicitly what applied researchers seem to assume would be the case if instruments are irrelevant. The reason is evident in Figure 6d. 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 from section I, when we find a spurious correlation in the reduced form, endogeneity between conflict and aid creates cointegration 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.

Figure 7: Monte Carlo panel IV estimates with negatively co-trending variables

7a: first stage estimate distribution *7b: reduced form estimate distribution*

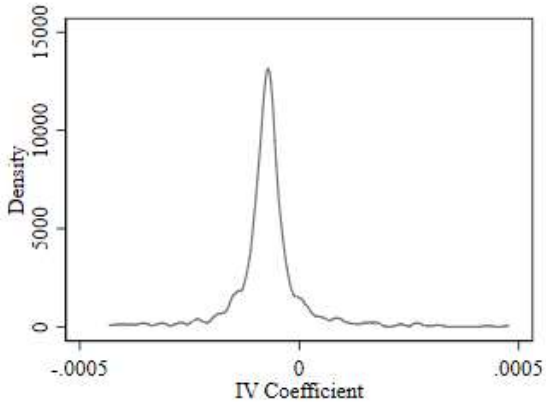
[21] The concern about convergence is not a negligible worry. As Yule (1926, pp. 12-13) put it "Be it remembered, we have taken a fairly long sample [to establish the independence of two cycling variables]… if the complete period were something exceeding, say, 500 years, it is seldom that we would have such a sample at our disposal." In other words, 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 A for a more detailed exploration of this issue.





7c: 2SLS coefficient estimate distribution

7d: Reduced form and first stage estimates



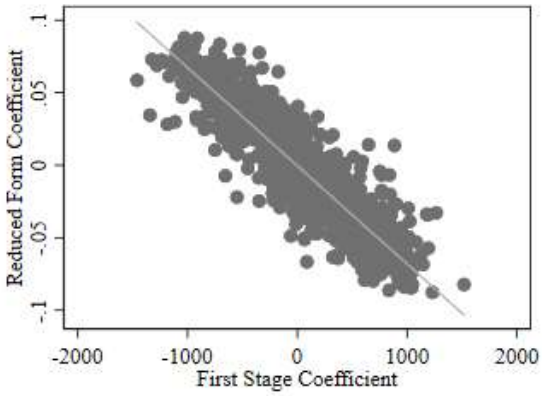


Figure 7 repeats the exercise, now using GDP growth rather than food aid as the endogenous X variable, following HI. The distribution of reduced form coefficient estimates in Figure 6a again shows the now-expected bimodal pattern with a disproportionate incidence of coefficient estimates farther from 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 7b that we saw in Figure 6b.

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 nonsense correlations, and the two may be correlated, as we saw in Figures 6 and 7, 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 in large enough samples, E[$\widehat{\gamma^{sim}}$ /$\widehat{\pi^{sim}}$]≠ 0 unless $\widehat{\pi^{sim}}$ and $\widehat{\gamma^{sim}}$ are uncorrelated. This problem is not uniformly related to the size of the sample. In simulations on increasingly long segments of the data, we find that the average of coefficients does not uniformly increase or decrease with larger samples. WeWeT

d. Addressing the common cycles problem

The fundamental issue with inference and identification in panel IV estimation is the strong assumption of iid errors, 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 (suppressing controls):

$$Conflict_{it} = \gamma_1^{ldv} Wheat_{t-1} + \gamma_2 Conflict_{it-1} + \mu_{it} \quad (26)$$

$$Aid_{it} = \pi_1^{ldv} Wheat_{t-1} + \pi_2 Conflict_{it-1} + \eta_{it} \quad (27)$$

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 are also nonstationary, as appears true in Figure 5a (and formal tests corroborate below), 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, spurious instrumental variable follows an I(1) random walk 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.

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. As is evident in Figure 8, 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 render errors iid when they exactly match the underlying time series process of the variables, an unlikely event. 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.

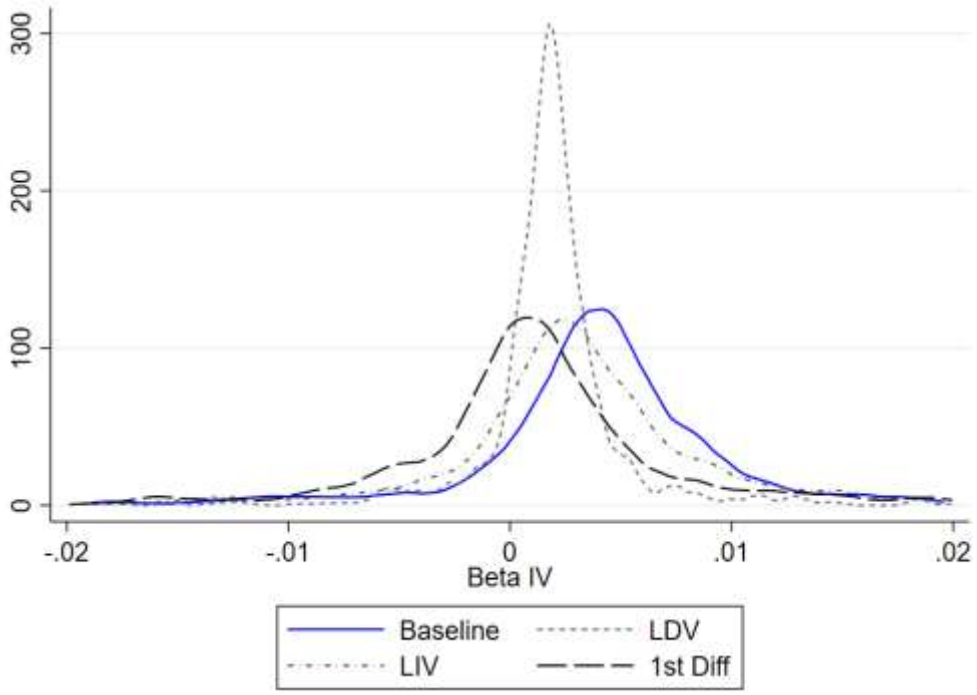
As reflected in Figure 8, the only specification that does not on average return an estimated positive effect of aid on conflict is the first differences regressions that exactly corrects for the known I(1) process of the manufactured instrument (controls again suppressed):

$$\Delta Conflict_{it} = \gamma_1^{diff}\Delta Z_t + \mu_{it}^{diff} \quad (28)$$

$$\Delta Aid_{it} = \pi_1^{diff}\Delta Z_t + \eta_{it}^{diff} \quad (29)$$

This works because it directly corrects for the known underlying nonstationarity of the time series variables.

Figure 8: Distributions of 2SLS parameter estimates



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 (13)-(14). 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. 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 from this section 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. In the single

equation case familiar from time series econometrics, the (reduced form or first stage) parameter estimates remain consistent, but classical inference tests overstate rejection rates on account of the multi-modal distribution of the parameter estimates. But in the case of IV estimation, the likely correlation between the reduced form and first stage regression estimates leads to bias and inconsistency. Ironically, in the case where reverse causality is a concern, the direction of bias in the IV estimate will match the sign of the reverse causal correlation between the dependent variable and the endogenous regressor, which the IV estimation was meant to overcome. Cycles of unknown periodicity make it difficult to control for trends effectively using lagged dependent or independent variables as controls. Corrections for trend or difference stationarity based on formal tests to identify the order of integration may work somewhat better, but precisely identifying the underlying time series process is a bit of a blunt instrument.

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

VARIABLES	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for intrastate war in year t - dummy for war in year (t-1)	Dummy for interstate war in year t - dummy for war in year (t-1)
ΔAid_t	-0.00802	-0.01155	-0.00832	-0.00726	-0.06586	-0.07786	-0.01625
	(0.01317)	(0.02421)	(0.01467)	(0.00938)	(0.49312)	(0.58185	(0.11351)
Controls (for all panels):							
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-year linear trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes
US real per capita GDP							
× avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
US democratic president							
× avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Oil price × 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 × avg. prob.							
of any US food aid	No	No	Yes	Yes	Yes	Yes	Yes
Avg. US military aid × year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. US economic aid							
(net of food aid) × year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. recipient cereal imports							
× year FE	No	No	No	No	Yes	Yes	Yes
Avg. recipient cereal							
production × year FE	No	No	No	No	Yes	Yes	Yes

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. 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 5: First-differenced 2SLS coefficients of food on conflict

VARIABLES	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for war in year t - dummy for war in year (t-1)	Dummy for intrastate war in year t - dummy for war in year (t-1)	Dummy for interstate war in year t - dummy for war in year (t-1)
ΔLn(GDP)$_t$	-13.02698	27.29817	-663.5553	-5.493698	5.035685	-4.358483	5.630787
	52.06428	179.9781	-.0014869	8.914968	5.492107	5.085462	5.105335
Controls (for all panels):							
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-year linear trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes
US real per capita GDP							
× avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
US democratic president							
× avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes	Yes	Yes
Oil price × 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 × avg. prob.							
of any US food aid	No	No	Yes	Yes	Yes	Yes	Yes
Avg. US military aid × year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. US economic aid							
(net of food aid) × year FE	No	No	No	Yes	Yes	Yes	Yes
Avg. recipient cereal imports							
× year FE	No	No	No	No	Yes	Yes	Yes
Avg. recipient cereal production							
× year FE	No	No	No	No	Yes	Yes	Yes

Notes: This table replicates 2SLS estimates of the effect of GDP growth on conflict using the HI approach of instrumenting for GDP growth with real interest rates, but with both conflict and real interest rates first differenced over time so that the instrument is real interest rates in year t minus real interest rates in year t-1. To make the table comparable to table 3, we use the same specifications and controls as in the NQ table. ΔLn(GDP)$_t$ is the growth rate in GDP, measured as the natural log of GDP in year t minus the natural log of GDP in year t-1.

III. 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 shift-share Bartik style 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:

$$Conflict_{it} = \gamma^{int} Z_t * D_i + \theta_i^{int} + t_t + \mu_{it}^{int} \quad (30)$$

$$X_{it} = \pi^{int} Z_t * D_i + \Theta_i^{\text{int}} + T_t + \eta_{it}^{int} \quad (31)$$

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 2 and 3), 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 DD estimator. When D_i is continuous this resembles a dose-response estimator.

Second, the instrument is interacted in both the reduced form equation (15) and the first stage equation (16). This allows for more flexible, nonparametric

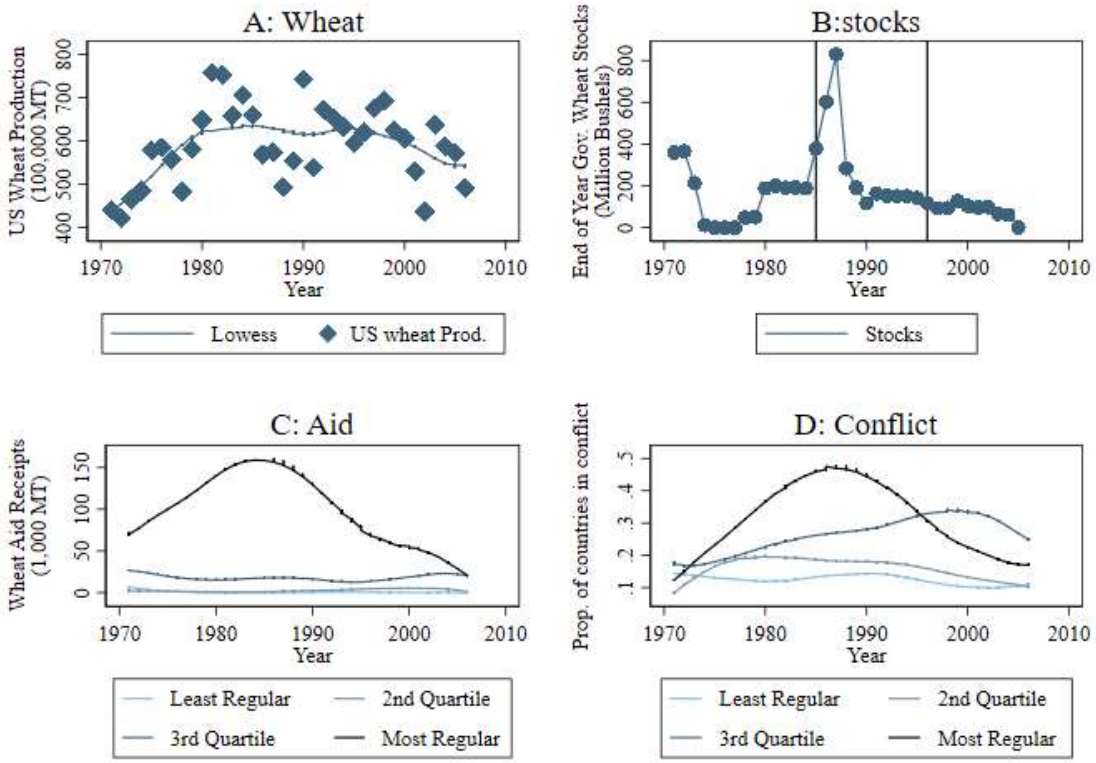
accommodation of unknown common trends, where t_t and T_t are a dummy variables indicating the year instead of the linear time trend *t* included in equations 2 and 3 with the uninteracted instrument.

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.[22]

The simplest way to see how the interaction strategy arises in the NQ setup is to plot the temporal variation of the key NQ variables and separate these trends on the same dimension as the interaction strategy. Figures 9A and 9B 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 9c visualizes the NQ shift-share identification strategy, 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 9d 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.

[22] Jaeger et al. (forthcoming, 2019) 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.

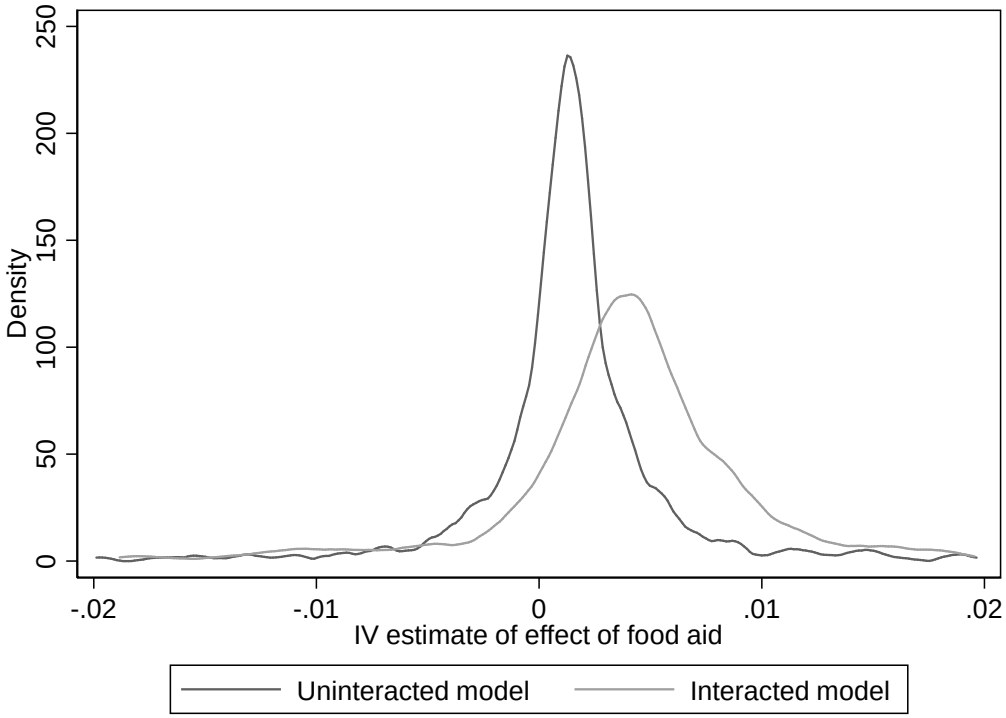
Figure 9: Time trends in the NQ variables



We now introduce the interaction term into the Monte Carlo simulation setup used in section II 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. Figure 10 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. Controlling for the flexible time trend does not remove the bias, because the distribution of coefficients in the interacted case is still centered above zero; they are merely rescaled by the interaction variable. Just as in the uninteracted case, using a spurious, non-stationary time series variable 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

Figure 11: Simulated distribution of 2SLS estimates using shift-share instrument



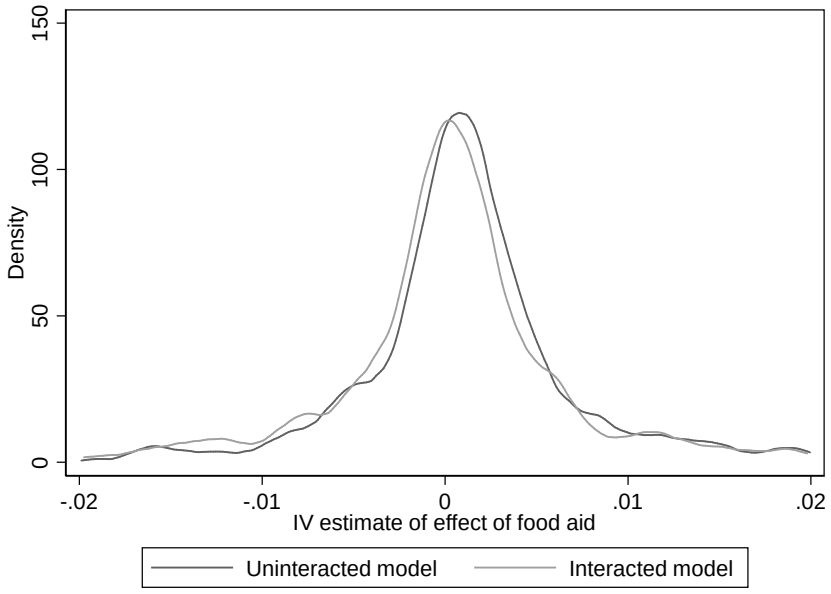
causality because of the spurious regression problem. Statistical power differs across the two samples, with a tighter distribution of coefficients in the interacted case than the uninteracted case, as Goldsmith-Pinkham (2018) describe can occur. This arises in the NQ case – and in Bartik instruments more generally – because the D_i term is defined on the [0,1] interval. Rescaling towards zero without removing bias necessarily concentrates the sampling distribution of the parameter estimates. Of course, that can make spurious rejection of the no-impact null hypothesis even more likely.

To understand how interactions affect the reliability of weak instrument tests, we test in our simulations how well weak instruments 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 while only 2.2% of irrelevant instruments pass this test in uninteracted models, 5.4% pass these tests in either the interacted or uninteracted models, meaning that allowing for weak

instrument tests to apply to either model means reducing power further. Because F-tests are weakly correlated across models, allowing for alternative interactions would reduce power of this tests still further. We describe these results in Appendix II, 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.[23] Figure 12 shows the distribution of estimated 2SLS coefficients of food aid on conflict. As was true in the uninteracted case, first differencing to remove spurious correlation in time series indeed eliminates the bias. Once one renders the errors iid, bias and inconsistency are gone. And because the interaction term now rescales without a bias constraint, it has no effect on the power of the estimator.

Figure 12: 2SLS estimates using shift-share instrument and first-differenced variables



[23] 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.

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 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 interaction term is in the unit interval, it thereby increases the statistical power of a biased estimator but does not remove the bias, which can exaggerate the risk of over-rejection of the null. 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 our example, by first differencing I(1) variables where the instrument is known to follow an I(1) random walk process. And then the interaction adds no statistical power to the uninteracted regressions. But it does retain 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 DD or dose-response estimators.

IV. Seven Practical Diagnostic Steps For Panel IV Estimation

The preceding cautions notwithstanding, in some cases panel IV estimation may still present the best option for attempting causal identification of a relationship of interest. But one must address several systematic concerns so as to increase confidence that panel IV results indeed yield unbiased estimates and to avoid the mistaken inference and bias that arise due to spurious regressions. Toward that end, we briefly outline seven practical steps one might take to try to systematically rule out the spurious regressions (and other) problems. Some will seem painfully obvious. Yet we observe that many published papers omit, and seem to fall prey to, these simple diagnostics. We number the steps merely as a

means of differentiating them, not to imply a necessary sequencing of steps. We illustrate this with application to the celebrated NQ paper, with details found in Appendix III.

Step 1: Carefully visually inspect the data

The first step in any empirical study should be to visually inspect the data for patterns. In the panel IV case, one needs to look not only at the scatter plots, which implicitly treat observations as iid, but also at the underlying time series in the outcome variable, the endogenous regressor(s), and the candidate instrument(s). Figures 2, 3, and 9 illustrate this. Recognizing that in the presence of cycles of unknown periodicity, one is in effect identifying off of clusters of sequenced observations, there is also merit in looking at trends in sub-periods of the sample. This can help pin down the source(s) of variation used to identify the relationship of interest. In the NQ case, as we illustrate in Appendix III, plotting simple trends at decadal scale in the sample shows that identification comes from a positive association between lagged wheat production and conflict that only differs significantly between regular and irregular recipients in the 1970s. When identification comes from a specific sub-period in the data, the analyst might usefully reflect on both (i) the representativeness of that sub-period for the broader inferences being made, and (ii) prospective confounders that might generate spurious correlation during that sub-period alone or a true causal mechanism that existed only during that sub-period.

Step 2: Identify most likely direction of OLS bias, then interrogate 2SLS estimates

One should always start regression analysis with simple ordinary least squares (OLS) estimation. Then carefully think through the mechanism(s) by which non-random selection, reverse causality, or some other form of statistical endogeneity would bias the OLS coefficient estimate(s). What direction of bias does this suggest in the OLS estimates?

To address the prospective bias in the OLS estimates, one will naturally attempt IV estimation. But it is essential to think carefully about two things. First, in the case of interacted IVs, does the cross-sectional component satisfy the exclusion restriction, or might it be endogenous, in particular vulnerable to reverse causality? More generally, one needs to compare the resulting 2SLS estimates with the previous OLS estimates in order to assess whether the resulting change in the coefficient estimate of interest is consistent

with a plausible mechanism that would bias the OLS estimate. If the IV estimate moves in a direction other than what one anticipated, that should serve as a caution. Perhaps one's understanding of the source(s) of bias in the OLS regression was flawed, and some other, countervailing mechanism overpowers the source of bias originally hypothesized. That explanation suggests a need to study the underlying mechanisms more carefully to establish whether such a countervailing mechanism is plausible. Alternatively, the IV results are spurious. As we demonstrated above in the case of reverse causality, the IV estimates biased by spurious regressions will move in the wrong direction relative to the OLS estimates.

NQ find a negative but statistically insignificant OLS association between US food aid shipments and conflict in recipient countries. NQ justifiably worry that this relationship could be biased, however, because food aid deliveries may be endogenous to conflict incidence. Since stated US government policy directs food aid flows to conflict affected countries, the bias in the OLS estimates should be upward. The fact that US food aid flows disproportionately to food aid affected countries is easily corroborated in the data. The fact that the 2SLS estimates increase relative to the OLS ones suggests either spurious regressions, or an odd, rather implausible negative selection mechanism wherein logistical, safety and other concerns about shipping food aid to conflict-affected countries overrides published policy. No evidence is provided to support the claim of negative selection. Furthermore, as we show in Appendix III, there is strong reason to believe that the shift-share instrument they use is endogenous, with reverse causality between the dependent variable and endogenous regressor leading to spurious regression bias in the 2SLS estimates that reinforces, rather than corrects for, the OLS bias, following precisely the mechanism we modeled in section I.

Step 3: Identify causal mechanism(s) and instrument(s); if possible, run placebo tests

Given the prospect of spurious correlation, one must clearly identify one or more candidate causal mechanism(s) and instrument(s) that would both directly reflect the mechanism(s) and pass the relevance and exclusion criteria for instruments. It is simply too easy to find variables that follow the same time series pattern as the endogenous

regressor and the dependent variable to trust an instrument that lacks a credible theory of the causal mechanism. Some variables that follow similar trends are obviously spurious, as we illustrated with the case of global audio cassette tape sales. But for many others, an analyst might be understandably tempted to inductively rationalize a mechanism that justifies a seemingly-relevant instrument. In the case of US food aid shipments, for example, we have already shown that real interest rates appear relevant as a proxy for US food aid shipments. One might imagine telling a story of how rising real interest rates depress consumer demand and raise the cost of holding grain stocks, thereby increasing food aid shipments. The danger of spurious regressions based on seemingly-relevant instruments is very real.

One way to test the plausibility of the causal mechanism one hypothesizes is to find a suitable placebo. In some cases, the placebo can arise from changes that occur within the sample period. Indeed, if during the step 1 temporal disaggregation of the data one notices distinct sub-periods within the data where the hypothesized relationship holds and others where it does not, a natural candidate placebo test arises from any change(s) that might have occurred to the hypothesized mechanism during the sample period. Alternatively, there might be within-sample cross-sectional variation – other than in the dimension of the interacted instrument, D_i – such that one sub-sample should exhibit the hypothesized mechanism and the other should not. Many data sets and hypothesized causal mechanisms thereby offer natural placebo tests through sample splitting, if one just studies the mechanism to identify prospective separatrices.

Let us illustrate with reference to NQ, which offers a very plausible mechanism to link the endogenous regressor, US food aid shipments, to the chosen instrument, lagged US wheat production. For many years, US farm policies effectively obliged the US Department of Agriculture (USDA) to purchase wheat in high production years when prices fell. NQ argue that the addition to USDA-held wheat stocks translated into extra wheat food aid sent mainly to the most regular US food aid recipients. Those claims are demonstrably true and previously established (Barrett and Maxwell 2005). NQ then cleverly exploit the resulting difference in additional food aid allocations between high and

low US wheat production years across regular and irregular food aid recipients to try to identify a causal effect of food aid flows on conflict in recipient countries.

NQ failed to acknowledge, however, a crucial change in US farm support and food aid procurement policies over the sample period (Barrett and Maxwell, 2005). Most notably, the wheat price support policies that provide the exogenous mechanism linking lagged US wheat production to food aid shipments changed, starting with the 1985 Farm Bill and culminating in the 1996 Farm Bill, which formally uncoupled government purchases from price or production targets (Willis and O'Brien, 2015).[24] This policy change implies that NQ's hypothesized first stage relationship between US wheat production and food aid deliveries should be strongest prior to 1985 and should disappear or at least attenuate after 1996. Note that variation by sub-period would be consistent with the period-differentiated effects we observed in step 1.

The latter portion of the sample period therefore offers a natural placebo test as the US government wheat procurement mechanism tapered off after 1985. But, as shown in detail in Appendix III, however, the placebo test fails when we re-estimate for sub-samples before the 1985 policy shift announcement, from 1985-1996, and after the wheat price stabilization policy completely ended in 1996. Rather than finding the hypothesized positive first and second stage relationship in the pre-1985 period, and no correlation in the latter period, we find a positive association both between food aid deliveries and wheat production (the first stage) and between conflict and instrumented food aid (the 2SLS estimates) in both the early (pre-1985) and latter (post-1996) periods. The pre-1985 estimates when the mechanism was mostly strongly in effect are statistically indistinguishable from the post-1996 ones when the hypothesized mechanism no longer

[24] Following the 1996 Farm Bill, the USDA was still enabled, though not obligated, to purchase wheat through the Commodities Credit Corporation. It is possible that wheat yields would remain associated with price or supply changes that could differentially create incentives to purchase wheat for food aid shipments under the Bill Emerson Humanitarian Trust or 416(b) programs in years with elevated wheat production, rather than draw on government-held stocks. However, such episodes have been very infrequent and when they have occurred, the bulk of the wheat procured has been used primarily for non-emergency Title II shipments that are monetized by the recipient NGO, i.e., not in emergency situations where food aid might prolong a conflict, as NQ hypothesize.

existed. This fails the placebo test that naturally arises from an exogenous change in the hypothesized mechanism during the sample period.

The US also operates multiple food aid programs, not each of which would be amenable to the surplus disposal mechanism NQ hypothesize. As explained in Appendix III, the evolution of US food aid shipments by program type correspond closely in time with the changes in farm policy. So effectively we ran a joint time series and cross-sectional placebo test. The failure to pass the placebo test meant to certify the plausibility of the hypothesized causal mechanism as the true source of identification should serve as a red flag for prospective spurious regression problems, since the statistical association exists in sub-samples where it should not.

Step 4: Test for trend and difference stationarity

Incorrect dynamic specification of the regression of interest is the source of the spurious regressions problem. So one needs to make a concerted effort to establish the most appropriate dynamic specification. The problem is that one never knows the true underlying ARIMA process underlying the time series. It is thus highly unlikely that standard, simple specifications –such as including a (potentially groupwise-) linear time trend or year fixed effects in the presence of interacted instrument – will render the error term iid. One needs to test directly for trend and difference stationarity in the underlying series and then correct for them, if/as necessary. As shown earlier, such correction resolves the spurious regressions problem.

The natural place to start is with non-stationarity tests of the dependent variable, the endogenous regressor, and the instrument series to check for difference and/or trend stationarity. Then difference and/or de-trend series as appropriate to the test results. Then estimate the OLS and IV models having appropriately corrected for the time series properties of the underlying variable. When the instrument is an interaction between a plausibly exogenous time series variable and a potentially endogenous cross sectional variable, researchers might begin by conducting an augmented Dickey Fuller (ADF) test to the time series component. The interacted instrument, dependent variable, and endogenous variable will typically be panel variables, so panel versions of unit root tests

are needed for those. In Appendix III, we use a Fisher-type test relying on ADF tests of each panel in the NQ data to test the hypothesis that all panels have a unit root and a Hadri test of the null hypothesis that all panels are stationary. These tests routinely fail to reject the unit root null and do reject the stationarity null, respectively, suggesting that the time series properties that we flag as potential cause for concern in panel IV estimation indeed pose an issue in these data.

Complete correction for trend stationarity is difficult in practice. For the panel variables, when unit root tests indicate that all panels are non-stationary, differencing should be applied. But these tests will often indicate that at least one panel, but not all panels, contain a unit root. In such cases, no one-size-fits-all correction works. Researchers might then report multiple specifications with different trends, as we do in Appendix III. Because interacted instruments allow researchers more degrees of freedom in choosing time trends, reporting one specification with an F-statistic above 10 is not a guarantee that instruments are robust. It is therefore good practice to report F-statistics and Anderson-Rubin (AR) confidence intervals for the coefficient of interest under multiple specifications that control for different trend specifications (linear, quadratic, etc.).

NQ report results controlling for a linear trend, for lagged dependent variables, and using other conventional controls without testing for trend or difference stationarity in the underlying series. As we show in Appendix III, ADF tests fail to reject the null of difference stationary series. And as we demonstrated in Table 3, re-estimating their model using first-differenced series overturns their core findings. But even if one just includes a quadratic rather than a linear trend in their regressions, without first-differencing, their results evaporate. Any association NQ identify cannot be distinguished from any other that might exhibit a quadratic trend, including both random cycles arising from summed errors over years or omitted variables with quadratic trends.

Step 5: Run Young's bootstrap test for overleveraged observations

Young (2018) demonstrates how conventional tests of significance in IV can be wrong. The primary issue Young identifies is high leverage of a small number of observations or clusters of observations. He convincingly demonstrates this by re-

estimating IV regressions, dropping one cluster at a time and bootstrapping the resulting p-values. This seems a very sensible check to employ in any IV estimation.

Yet Young's procedure does not necessarily address the concern we describe for the simple reason that correlated trends are a different problem from leverage. In the limit, a time series with no variation around its trend has zero observation-specific leverage. Dropping one country (or one year) from the analysis does not generally change the estimation results because identification is coming from trends that are common across countries (and years). Our concern about spurious regressions in IV estimation thus complements Young's. We therefore favor including Young's diagnostic tests once one has corrected to the maximum extent possible for any difference or trend stationarity present in the time series so that the iid assumption (approximately) holds.

As we show in Appendix III, when we implement Young's bootstrap test on the NQ data, we find that the p-values of the coefficient estimates rise slightly, but remain significant at the five percent level. Over-leverage does not appear to be a first-order concern in that specific paper.

Step 6: Check the instrument using a randomization placebo test

Step 3 advocated the use of a placebo test to ensure that when the hypothesized causal mechanism is not in effect, the IV estimator fails to reject the null hypothesis of no effect. If the IV estimator still rejects the null when it should not, that signals that some other (spurious) source of correlation must be the source of the observed partial correlation between the instrumented regressor and the outcome of interest. In the continuous interacted (e.g., Bartik) instruments case, one way to run the placebo test is to preserve the meta-structure of the data, but randomize the endogenous regressor among observations at the intensive margin, generating a randomized 'pseudo-dose' for which there should be null response. More specifically, we randomly reassign non-zero observed values among cross-sectional units within a given time period, holding constant everything else in the data. This disables the hypothesized causal mechanism through randomization while preserving the two sources of bias that concern us: the potentially-endogenous shift-share instrument, and the time series structure of the instrument, (aggregated) endogenous

regressor, and outcome variable. If the IV estimator still rejects the zero correlation null, this would seem strong evidence that spurious regressions or of violation of the exogeneity/excludability criteria for instruments explain the original result, not the hypothesized causal mechanism. We randomize and estimate 1,000 replicates to generate the distribution of the placebo test parameter. As best as we can tell, this appears a novel randomization placebo test method.[25]

Appendix III describes in detail how we implement this novel randomization placebo test with the NQ data. This method preserves the annual aggregate US food aid flows, but reassigns actual values for a given year among countries that received food aid shipments that year. Recall that the hypothesized mechanism is that the aggregate food aid supply shock attributable to a US wheat production shock the preceding year propagates disproportionately to the most frequent food aid recipients. So we preserve the entire interacted instrument, including the (potentially endogenous) shift-share cross-sectional component, but match these with some other recipient country's food aid flow that year. Since the true shipment to, say, Guatemala, cannot possibly explain variation in conflict in, say, Bangladesh, this randomization generates a placebo test while preserving the whole sample size (and associated precision) and, importantly, any spurious regressions due to non-iid errors. If indeed food aid flows are the cause of the IV estimate on conflict through a mechanism based on differential exposure to the time series innovations in the instrument, then the randomization placebo test should eliminates the correlation, a placebo test. As we show in Appendix III, the placebo test still rejects the null, signaling that the correlation observed in the data does not arise due to the causal mechanism NQ claim. Indeed, we find the placebo test IV parameter estimate distribution shifts markedly upwards relative to the NQ estimate, signaling that any true causal effect food aid might have on recipient country conflict must be negative, not positive.

[25] This randomization placebo test is somewhat similar to the randomization method Keller (1998) employs. We thank Joe Kaboski for pointing this out.

Step 7: Monte Carlo placebo tests: simulate after eliminating/reversing causal mechanism

The placebo tests of steps 3 and 6 may not fully convince the most skeptical analyst. In the step 3 case, it may be difficult to find an appropriate sub-period or sub-sample to use; or it may just be hard to know the true null within the appropriate sub-period or sub-sample. In the step 6 case, the randomization method intentionally generates a weak instrument, which could account for null results. A more complicated but reasonably foolproof method to placebo test one's estimation results is to build a simple, stylized model to replicate the mechanisms one believes drive the associations seen in data, and then to mechanically break, or even reverse, the mechanism in the model. Monte Carlo simulation using the same data one uses in estimation but with a known data generating process (DGP) that eliminates (reverses) the causal mechanism should generate null (opposing) results. This eliminates any uncertainty about the null under the placebo test and the weak instruments problem.

Appendix III details the Monte Carlo model we build to replicate the relationship NQ hypothesize between US food aid flows and conflict, except that we expressly eliminate the causal link. Then we introduce a DGP in which food aid flows stochastically reduce the incidence of conflict, reversing the causal mechanism relative to NQ's hypothesis. When we then simulate coefficient estimates using the NQ data, we find that even when food aid flows are statistically independent of conflict in the true DGP, the NQ estimation strategy yield estimates spuriously suggesting a causal effect of food aid flows on conflict. The estimation strategy fails the Monte Carlo placebo test. Indeed, even when we construct the DGP so that food aid flows causally reduce conflict, we show that spurious regressions will generate OLS and 2SLS parameter estimates qualitatively identical to NQ's. The takeaway message is powerful. We can replicate the NQ results even if US food aid agencies prefer to send food aid to conflict-affected countries – as is the stated policy of the US food aid program (but opposite to how NQ explain the sign shift between their OLS and 2SLS estimates) – and food aid has no causal effect on – or even *prevents* (not prolongs) – civil conflict in recipient countries. The key drivers of this result are (i) greater long-run variation than short-run variation in the exogenous component of the instrument

(US wheat production) and the outcome variable (conflict), combined with (ii) spurious correlation in those longer-run trends, as we already showed characterizes the underlying data. These results strongly suggest that NQ's findings arise from spurious regressions, not from a true causal effect of US food aid flows on recipient country conflict.

If one follows these seven diagnostic steps, identifying and correcting for any difference and trend stationarity in the time series, and passing Young's test and the various placebo tests we propose, then one could reasonably conclude that panel IV estimates are indeed causal. But if the panel IV results fail multiple of these diagnostic tests, as we show in Appendix III the NQ results do, then one should conclude that no true causal relationship existed.

V. 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. 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) instruments do not resolve that problem; they merely rescale it. Much like Bazzi and Clemens (2013), we offer a caution about instrument validity and strength in panel data IV estimation. We also offer seven practical diagnostic steps researchers can follow to minimize the risk of bias and mistaken inference due to spurious regressions in panel IV estimates. Applying these techniques to a celebrated recent paper that claims to find that US food aid shipments cause prolonged conflict in recipient countries, we find those results highly contestable.

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.

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 (2018), "Bartik Instruments: What, When, Why, and How", NBER working paper 24408.

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. "Recording Industry in Numbers." Accessed 26 Aug 2015. https://musicbusinessresearch.wordpress.com/2010/03/29/the-recession-in-the-music-industry-a-cause-analysis/.

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 (forthcoming). “A Cautionary Tale of Evaluating Identifying Assumptions: Did Reality TV Really Cause A Decline in Teenage Childbearing?” *Journal of Business and Economic Statistics*.

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/

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-Simulation results with increasing time dimension

In addition to biasedness, one may wish to interrogate the role of neglecting variables' time series properties on consistency. The above Monte Carlo approach can be used to investigate consistency by re-running the same simulations with increasingly long time series. Table A1 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 A1: 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 capture 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 B: Weak instruments tests in uninteracted models

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 reasonably 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.

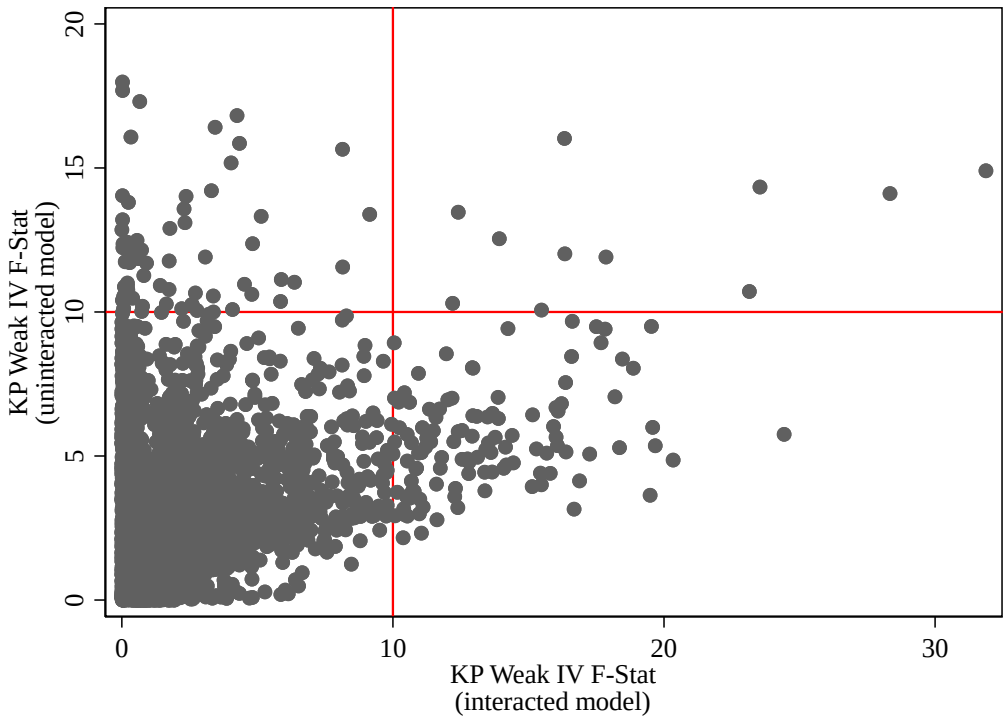
Introducing a shift variable to the instrument has problematic outcomes for the IV strategy. When the irrelevant variables are interacted with D_i, we now find that the weak instrument tests passes the threshold of 10 in 3.5% of the cases, meaning that the interaction model increases the likelihood that weak instrument tests falsely conclude that irrelevant instruments are strong by 63%.

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 are simply making good instruments stronger, or whether strong instruments appear through the influence of the interaction rather than the plausibly exogenous time series instrument. In Figure B-1, we show the scatter 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.

To understand how the noisiness of this correlation affects the practice of implementing weak instrument tests, consider implementing one of several possible rules

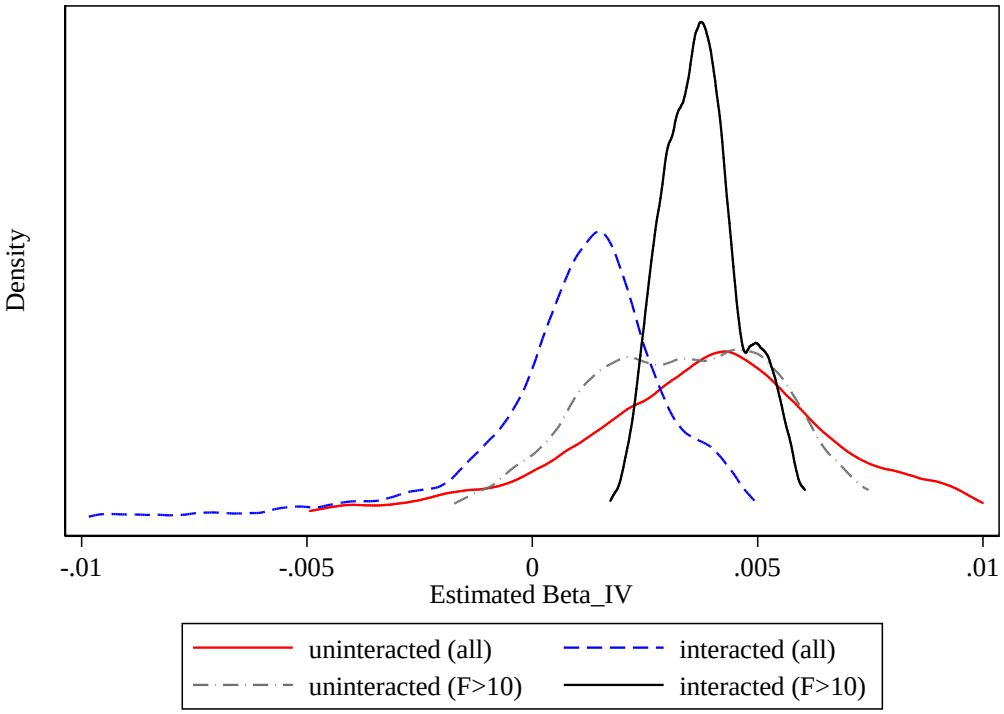
for whether or not to accept an IV as strong. First, suppose we accept instruments as valid only if they pass the weak IV test of F>10 in the uninteracted case. In our simulations, this would mean accepting as valid only 2.2% of the irrelevant instruments. Second, suppose we accept instruments as valid 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. Thus allowing for interacted specifications 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.

Figure B-1: Correlation of Weak IV Test Statistics for Interacted and Uninteracted Models



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 consequence of this approach is that we can never know how many endogenous cross-sectional variables were tried to find an F-statistic greater than 10, meaning we can never know the true power of weak instrument tests for the interacted models.

The other relevant consideration in the role of F-stats is whether using weak instrument tests help us avoid the bias that arises when we falsely accept an irrelevant instrument as valid. In the figure below, 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 weak instrument tests do not avoid the bias arising from irrelevant instruments. 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 line), we find that the distribution of IV coefficients among strong instruments only is more biased than the distribution of coefficients without strong instruments.



The conclusion is that allowing for interacted models to pass weak instrument tests not only does not solve the spurious correlation problem that is the core concern of this paper, it reduces the power of weak instrument tests and increases the bias among the spurious instruments which pass.

Appendix C: Applying the seven practical diagnostic steps to revisit NQ

In this appendix, we apply the seven practical diagnostic steps for panel IV estimation we advance in the main paper to revisit the celebrated Nunn and Qian (2014, hereafter NQ) study. We focus on the NQ results because they have been widely publicized and inform an intensely debated policy issue that is especially timely as the future of US foreign assistance and food aid in particular are under serious scrutiny in Washington. If a policy commonly labeled "humanitarian" actually causes violent conflict, that policy should be revisited. We show through a sequence of diagnostic tests that their headline finding – that US food aid shipments cause prolonged conflict in recipient countries – does

not pass several placebo tests, and appear to result not from the causal mechanism they posit but instead from spurious regressions arising from longer-run trends in nonstationary data series, combined with a shift-share instrument that fails the exogeneity criterion for valid instruments. Indeed, Monte Carlo simulations explained below demonstrate that the NQ results could actually arise from a data generating process in which food aid is independent of or even *reduces* conflict, contrary to their core claim.

We note that the results reported by NQ have also been disputed by USAID (2014), who report on robustness of the NQ strategy to controlling for other forms of non-food aid external support for actors in civil conflicts including use of external military bases and economic support for rebels. USAID suggests that when these variables are included as controls, the statistical significance of food aid disappears. Unfortunately, the USAID results are not directly comparable to the NQ strategy for two reasons. First, external support to combatants only occurs by definition when a conflict exists. Food aid, on the other hand is sent to both countries that are experiencing conflict and those that are not, so that the NQ dataset can leverage information from countries that are not actively experiencing conflict. Second, the external aid variable is not available for the earliest years of NQ's dataset. If NQ identify an association that is strongest in the early period, as we show below is the case, then the USAID estimation strategy could miss the effect by omitting those years. USAID argues that the NQ results are fragile with regard to these robustness checks, but they are not able to fully explain why the NQ strategy identifies a positive association between aid and conflict. The threats to identification we identify through the seven diagnostic test steps detailed in this appendix directly explains both why NQ found an effect, and why one should conclude that it is a spurious, not a causal, association.

Step 1: Carefully visually inspect the data

An intuitive way to show how such time trends influence NQ's IV estimate is to reproduce the plots NQ use to explain and demonstrate their strategy, highlighting changes across decades. Figure C1 reproduces the NQ's Figures 3 and 4, which show the relationship between US wheat production and the proportion of countries in each year

who are experiencing a conflict. The bottom panel shows the relationship among only the 50% of countries that received aid from the US most frequently – in at least 30% of periods – while the top panel is the 50% of countries that received aid least frequently during the study period. NQ use these plots to show that that wheat production is related to conflict, but only among frequent recipients of US aid, presenting an intuitively appealing demonstration for what is effectively the reduced form in their IV strategy.

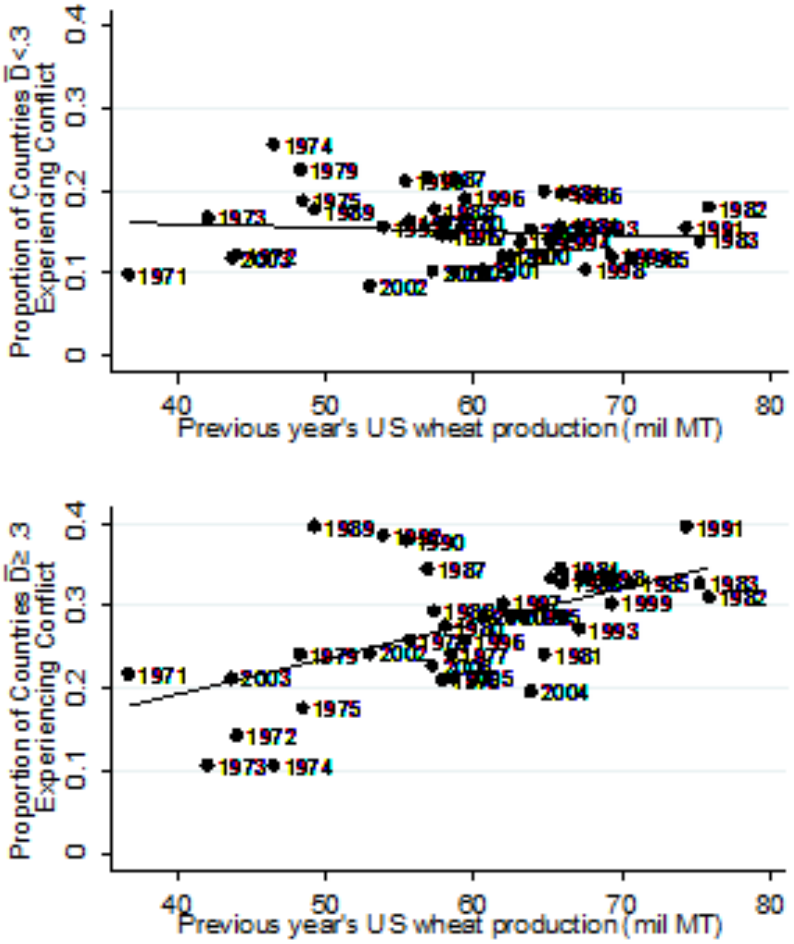


FIGURE C1: THE ESSENCE OF THE NUNN AND QIAN IV STRATEGY

Notes: The figure above replicates the core results from figures 3 and 4 in NQ. Because this paper focuses on the presence of any type of conflict as the headline result from the NQ study, the y-axis is proportion of countries experiencing any type of conflict rather than only intrastate conflicts as in the original NQ figures. The figures are qualitatively very similar if only intrastate conflicts are used.

But given that wheat production displays a pronounced nonlinear trend in the sample (Figure 3), it is useful to group observations that are near each other in time. Figure

C2 (A3) reproduces the preceding figure for the low (high) frequency food aid recipients, with different markers and colors representing different decades. As we know from the fact that both food aid and conflict followed the same inverted-U shape trend during this period as US lagged wheat production (Figure 3), all of the years with high wheat production and elevated incidence of conflict occur in the 1980s and 1990s (grey diamonds and dark blue squares, respectively), while all the years with low wheat production and high conflict occur in the 1970s and 2000s (light blue circles and black triangles, respectively).

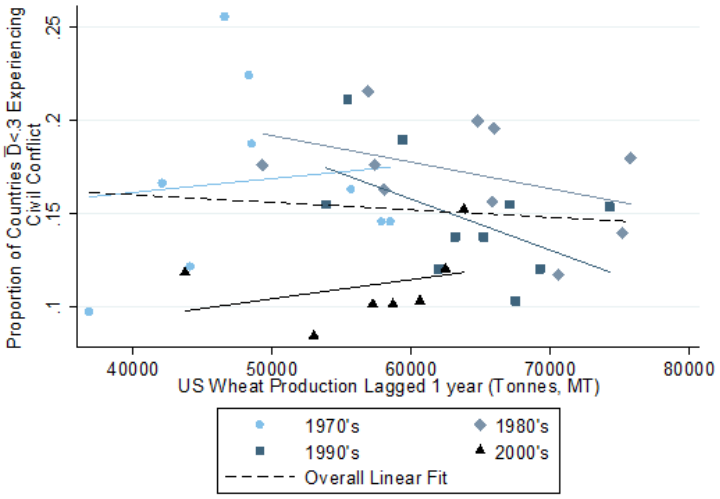


FIGURE C2: TRENDS IN CIVIL CONFLICT INCIDENCE FOR COUNTRIES WITH $\bar{D} < 0.3$ (IRREGULAR AID RECIPIENTS)

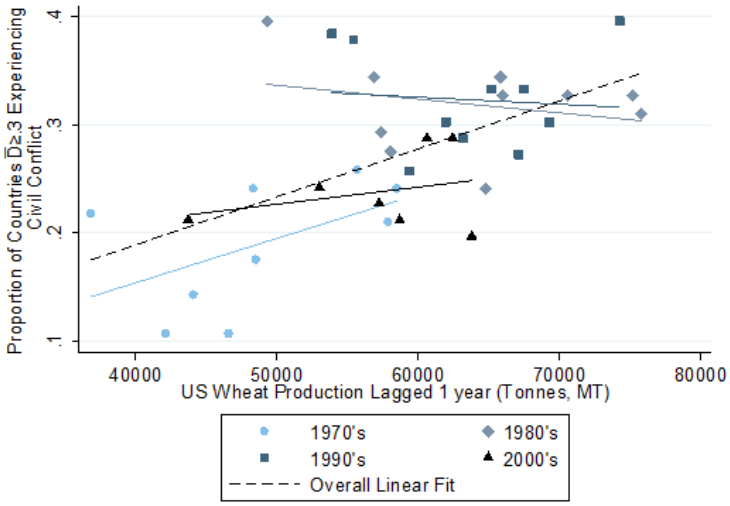


FIGURE C3: TRENDS IN CIVIL CONFLICT INCIDENCE FOR COUNTRIES WITH $\bar{D} \geq 0.3$ (REGULAR AID RECIPIENTS)

These time trends are important for NQ, because their instrument is based on the interaction of US wheat production and a country's long-term propensity to receive aid. Lagged wheat production drives the plausible exogeneity of the instrument, but it only varies by year. NQ's results depend on conflict increasing more among regular recipients than among irregular recipients when lagged US wheat production is high.

Figures C2 and C3 show what happens when depicting the reduced form relationship between lagged wheat production and conflict – the reduced form relationship of interest – for irregular and regular recipients separately by decade. In Figure C2, we see that if anything, within any given decade, food aid and conflict are negatively correlated among irregular food aid recipients. What had previously appeared to be a flat relationship in the top panel of Figure C1 was driven by the fact that US wheat production and global conflict were both higher in the 1980s and 1990s. Among the regular recipients, Figure C3 shows that food aid appears related to conflict in the 1970s and possibly the 2000s, but not at all in the 1980s and 1990s. What appeared to be a globally positive relationship between lagged wheat production and conflict in the bottom panel of Figure 1 is entirely driven by a transition in the late 1970s to a period of high US wheat production and high recipient country conflict and back to a period of lower US wheat production and global conflict over the 2000s. This transition corresponds to a period of dramatic shifts in US farm price policy (on which, more below) that broke the hypothesized narrative of a mandated link between wheat production and government-held wheat stocks and during which US food aid policy began expressly prioritizing the shipment of emergency food aid to conflict-affected countries.

What matters for the average relationship that NQ identify is the difference in the conflict-wheat relationship between regular versus irregular aid recipients. This is shown in Figure C4. Their hypothesized effect seems present in the 1970s, but otherwise, the global relationship is almost entirely driven by the long-term changes rather than the more plausibly random short-term fluctuations of wheat output around that trend, off which they

ostensibly identify causal effects. Moreover, if there is an upward relationship in the differences shown below for the 1990s, then it is entirely driven by the fact that average conflict was declining in irregular recipients when wheat production was low rather than the fact that it was increasing in regular recipients, among whom it was actually flat or declining, as shown in Figures C2-C4, respectively.

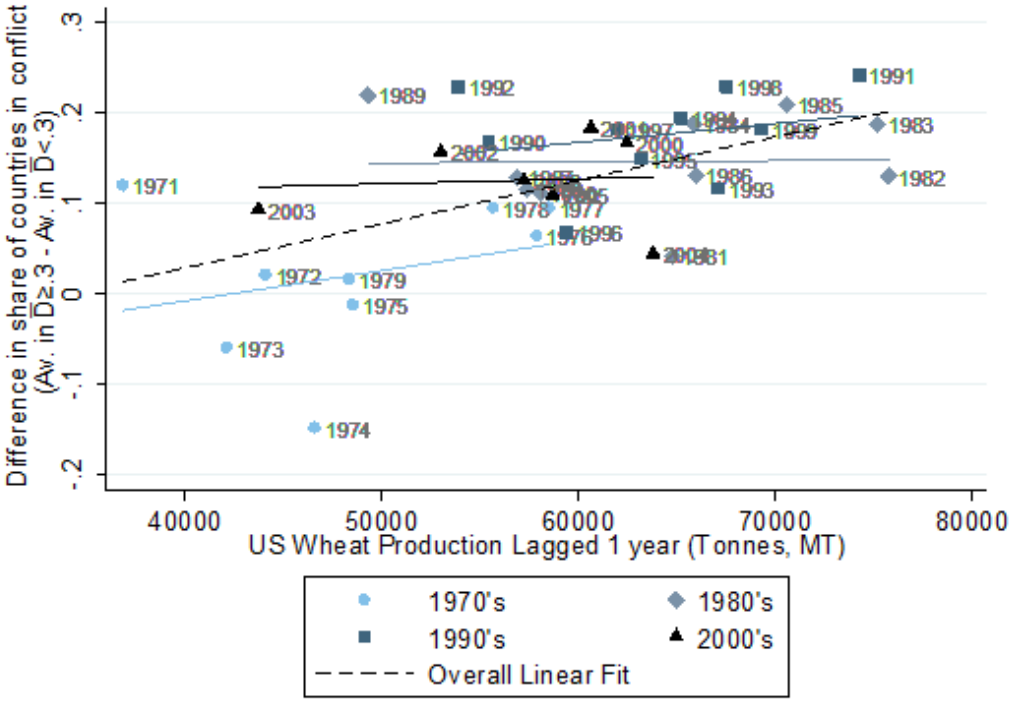


FIGURE C4: LINEAR TRENDS IN THE DIFFERENCE BETWEEN AVERAGE CIVIL CONFLICT INCIDENCE FOR COUNTRIES WITH $\bar{D} \geq 0.3$ AND COUNTRIES WITH $\bar{D} < 0.3$ (REGULAR MINUS IRREGULAR AID RECIPIENTS) BY DECADE

By itself, the coincident trends between lagged wheat production and conflict do not necessarily mean that NQ's IV strategy does not identify a causal effect of aid on conflict. But the visual inspection of the time series by sub-period makes clear the source(s) of identifying variation is limited to the 1970s. One possibility, discussed in more detail in

step 2 below, is that the driving causal mechanism was most salient only during that sub-period. But it is equally possible that since variation in wheat and conflict appears mostly driven by long decadal changes rather than year to year variation, there could well be omitted cyclical factors that are related to both wheat production and conflict. It is entirely possible that the evolution of two unrelated processes could correlate for a short period of time by coincidence, i.e., the spurious regression problem. Given that food aid is, by program design, intentionally directed toward countries most at risk of conflict, such coincidence would generate spurious correlation and bias in NQ's IV estimates.

Step 2: Run OLS regression, identify most likely direction of bias, then interrogate 2SLS estimates

NQ first estimate an OLS specification[26] where the indicator variable for the presence of conflict in a given country during a given year is regressed on the quantity of wheat food aid delivered to that country:

$$C_{irt} = \beta F_{irt} + \boldsymbol{X}_{irt}\Gamma + \phi_{rt} + \psi_{ir} + \nu_{irt} \quad \text{(C1)}$$

where C_{irt} is an indicator variable which equals one if country i in region r experiences at least 25 deaths from battle involving two parties in year t, F_{irt} is the endogenous quantity of wheat aid shipments to country i in year t, $\boldsymbol{X}_{irt}$ is a set of country and year controls, ϕ_{rt} is a set of region- specific year fixed effects, and ψ_{ir} is a set of country fixed effects.

The OLS estimates of equation B1 suggest that additional food aid is associated with a lower incidence of conflict; but this result is not statistically significant. This result implies that within countries, there is no greater risk of conflict in years when a greater quantity of wheat aid is received from the United States, and within years, countries that receive a greater quantity of wheat aid have no greater risk of conflict.

NQ justifiably worry that this relationship could be biased, however, because food aid deliveries may be endogenous to conflict incidence. Since stated US government policy directs food aid flows to conflict affected countries, the bias in the OLS estimates

[26] Here we repeat the notation exactly as it appears in the original NQ paper for comparability.

should be upward. The fact that US food aid flows disproportionately to food aid affected countries is easily corroborated in the data.

In NQ's case, so as to obtain a causal estimate of the effect of food aid on conflict, they estimate a 2SLS specification:

$$F_{irt} = \alpha(P_{t-1} * \bar{D}_{ir}) + \boldsymbol{X}_{irt}\Gamma + \phi_{rt} + \psi_{ir} + E_{irt} \quad \text{(C2)}$$

Equation B1 remains the equation of interest. The first stage, shown in equation B2, interacts P_{t-1}, annual US wheat production lagged by one year,[27] and $\bar{D}_{ir}$, the proportion of the 36 years in which country i received a non-zero quantity of wheat aid from the US. NQ exploit the plausible exogeneity of P_{t-1} to identify the key parameter of interest, β, in the second stage.

The justification for this first stage is that when additional wheat is available because of high production (P_{t-1}), additional aid is sent to regular food aid recipients, which goes disproportionately to the favored aid partners. Including the country and year fixed effects, α is then analogous to a continuous difference-in-differences (DD) coefficient estimate, where the variation in the instrument comes from comparing aid between high and low wheat production years and between regular and irregular aid recipients. Any confounding variables that have a common effect on conflict across all countries within a region in the same year, such as weather or climate or global market prices, or characteristics of countries that have a constant effect on conflict prevalence over time are controlled for through region-year and country fixed effects.

In their preferred specification, NQ estimate equation B1 via 2SLS with the interacted instrument in first stage equation B2 used for endogenous food aid quantity, and find that "a 1,000 MT increase in US wheat aid increases the incidence of conflict by .3 percentage points." At the sample means, their estimated food aid elasticity of conflict incidence is 0.4, a large enough magnitude to warrant serious policy attention.

NQ's 2SLS conclusions depend crucially on the credibility of the exclusion restriction behind their IV strategy, which is that lagged US wheat production conditional

[27] Lagged one year due to the time required to plan and implement food aid deliveries.

on the set of controls reported in the paper is correlated with conflict *only* through the following channel. Positive shocks from higher wheat production year-on-year lead the US government to purchase more wheat, which leads food aid agencies to distribute more wheat the following year. The resulting extra food aid is shipped primarily to frequent food aid recipients where pre-existing administrative and logistical arrangements make variation in food aid volumes at the intensive margin more feasible. These added food aid receipts cause conflict prevalence to increase, in particular by prolonging the duration of conflict.[28]

The first concern with the NQ IV strategy is the prospective endogeneity of $\overline{D}_{ir}$ which, after all, includes food aid receipts from *all* years in the sample. Goldsmith-Pinkham et al. (2018) offer a detailed technical exploration of this issue in shift-share instruments, to which we direct interested readers. In the present context, the crucial thing to understand is that US food aid shipments exhibit strong persistence over time (Barrett, 1998; Jayne et al, 2002; Barrett and Heisey, 2002). Barrett (1998) shows that historically the probability of a country receiving future food aid flows conditional on past food aid receipt is 85% or greater out to horizons of 35 years of prior food aid receipt. A shock, like conflict, that sparks the initiation of emergency food aid flows is therefore likely to have persistent effects on food aid flows.

Note that NQ find no evidence that food aid causes conflict to begin. Their claim is that food aid flows cause prolongation of a conflict. But if initiation of conflict is independent of food aid flows, as NQ find, and there exists reverse causality, per US government policy, then persistence in food aid flows can render $\overline{D}_{ir}$ endogenous and vulnerable to the Goldsmith-Pinkham et al. (2018) critique.

Figure C5 demonstrates the persistence of food aid flows in the particular years and countries of the NQ dataset by showing the distribution of years of duration of aid spells, defined as a continuous period over which a country receives aid in each year. We

[28] The exact channel by which this last stage occurs is not clear. NQ argue that the extra food aid is vulnerable to theft, which allows rebels to continue fighting longer than they otherwise would. But this seems to contradict findings in the literature that show that exogenous sources of income growth are often associated with lower conflict (Blattman and Miguel 2010, Ray and Esteban 2017). The mechanisms that would explain why income from aid deliveries would be more likely to be stolen by rebels and increase conflict while income from other sources often reduces fighting have not been established.

separately plot the distribution of spell lengths where there was a conflict in the country in the first year of the spell from those where the country was not in conflict when the spell began. Figure C5 clearly shows that once food aid starts flowing, the country is likely to continue receiving aid for many subsequent years. This is especially true if the recipient country experiences conflict in the year when the spell starts. This highlights the fact that aid allocations are highly endogenous; once aid starts flowing to a country, it is likely to continue, and this is especially true in conflict situations.

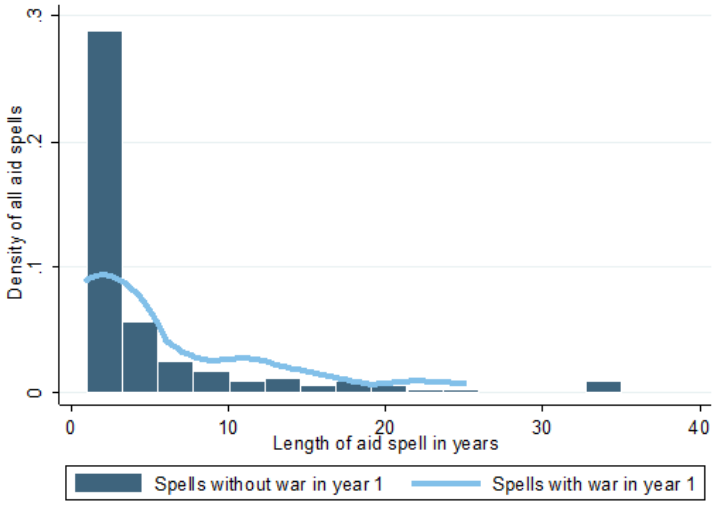


FIGURE C5: LENGTH OF AID SPELLS IN YEARS BY INITIAL CONFLICT STATUS

Notes: An aid spell is the number of uninterrupted years between observing a country receiving any wheat aid following either a year with no wheat aid or the start of the dataset, and the first subsequent year in which that country does not receive aid. The histogram shows the density of aid spells for countries not experiencing a conflict in the first year of the dataset. The overlaid kernel density plot shows the density of all aid spells for countries that did have a conflict in in the starting spell.

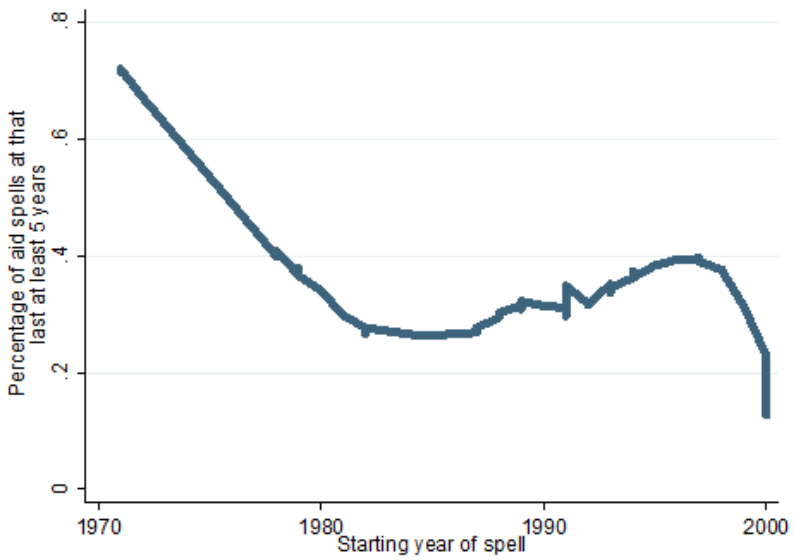


FIGURE C6: STARTING YEARS OF LONG AID SPELLS

Notes: Lowess plot of indicator for whether an aid spell lasted ≥5 years, by the starting year of the spell.

The degree of persistence has also changed over time as aid allocation priorities have changed. Figure C6 shows the percentage of aid spells that last at least five years, conditional on the year in which the spell started. Wheat aid was significantly more persistent in the 1970s when most food aid flowed as (Title I PL480) annual concessional exports to governments with established Title I programs. This persistence lessened in the 1980s and 1990s as Title II grants to NGOs and WFP began to replace Title I, catering to a different set of countries. The persistence grew stronger again in the 2000s when USAID began concentrating non-emergency Title II flows on just a few countries that routinely had emergency Title II flows for two reasons. First, USAID sought to use non-emergency assistance to preempt the need for emergency food aid flows. Second, having non-emergency food aid distribution pipelines in place and operational already conveys considerable administrative and logistical advantages in mounting effective and rapid emergency response when a disaster strikes or a conflict erupts (Barrett and Maxwell, 2005).

A second major concern is NQ's claim that the modest, statistically insignificant negative association between food aid and conflict in their OLS estimates suffers

significant *negative* bias. The NQ explanation for that claim is the possibility that donors condition food aid flows on characteristics correlated with low levels of conflict, i.e., that the US actively seeks to avoid sending foreign aid to conflict-affected countries. This explanation directly contradicts USAID's Food for Peace program's stated objective "FFP provides emergency food assistance to those affected by *conflict* and natural disasters and provides development food assistance to address the underlying causes of hunger." (USAID, 2015, emphasis added) In the NQ dataset, less than 22% of countries experience conflict in the average year, and yet between 1975 and 2006, there were only two years when there were more countries receiving food aid and not experiencing conflict than countries who were both receiving food aid and experiencing conflict. So their explanation does not appear to square with the data.

Indeed, as US food aid under PL480 shifted from Title I in-kind concessional lending to governments to Title II emergency assistance through PVO/NGOs and WFP, food aid has grown increasingly concentrated on populations dealing with conflict, contrary to the NQ hypothesis. If food aid deliveries intentionally directed toward countries that have conflict rather than away from such countries, reconciling a positive IV coefficient with a negative OLS coefficient becomes difficult. If humanitarian assistance during conflict is a primary source of endogeneity, one would have expected the OLS coefficient to be upward biased rather than downward as would be implied by NQ's reported effects.

The fact that the 2SLS estimates increase relative to the OLS ones suggests either spurious regressions, or an odd, rather implausible negative selection mechanism wherein logistical, safety and other concerns about shipping food aid to conflict-affected countries overrides published policy. No evidence is provided to support the claim of negative selection. And the shift-share instrument NQ does not seem to satisfy the exclusion restriction, being vulnerable to reverse causality and thus the Goldsmith-Pinkham et al. (2018) critique and the bias we demonstrate arises in panel IV estimation when spurious regressions occur in conjunction with reverse causality between the dependent variable and the endogenous regressor for which one instruments.

Step 3: Identify candidate causal mechanisms and instrument(s); if possible, run placebo tests

In NQ, the mechanism driving variation comes from US government wheat price stabilization policies – established for exogenous reasons entirely unrelated to US food aid programs (Barrett and Maxwell 2005) – that for many years obliged the USDA to purchase food in high production years when commodity prices fell. NQ argue that windfall wheat production resulted in extra government-held wheat stocks which were subsequently shipped abroad as food aid, with year-on-year perturbations absorbed by food aid shipments supporting countries that are the most regular recipients of US food aid. NQ then sensibly exploit the resulting difference in additional food aid allocations between high and low US wheat production years across regular and irregular food aid recipients to try to identify a causal effect of food aid on conflict in recipient countries.

The US long had a policy of agricultural commodity price supports that indeed created a link between aggregate annual production and government procurement of wheat that was subsequently used as food aid; indeed, surplus disposal was an explicit policy objective of the main US food aid program launched in 1954 (Barrett and Maxwell, 2005). But US commodity price stabilization and food aid policies experienced dramatic changes during the sample period NQ study. And the US operates several different food aid programs with important differences in features that make some more or less relevant to the causal mechanism NQ propose. These intertemporal changes and inter-program differences create opportunities to run placebo tests of the hypothesized causal mechanism.

In practice, policies that link production to US government procurement were not in place for the entire duration of the NQ study period. In the 1970s and early 1980s, the start of the NQ time series, purchases took place through USDA's system of non-recourse loans, which were essentially loans that the USDA made to US farmers through the USDA's Commodity Credit Corporation (CCC).[29] The USDA would purchase a farmer's

[29] A farmer could take out a non-recourse commodity loan proportionate to his or her harvested quantity of wheat at a fixed unit rate with the grain held as collateral. Within a nine-month window, if the selling price of grain dipped below the loan repayment rate, the farmer could forfeit the grain rather than repay the loan.

grain production at a fixed rate if the market price fell below that rate. In order to avoid having these large reserves putting downward pressure on future grain prices, the USDA donated commodity stocks to countries beyond its commercial marketshed as food aid through Section 416(b) of the Agricultural Act of 1949, and subsequently through the food aid programs authorized under Public Law 480 (PL480), passed in 1954, which thereafter became – and continues today as – the principal vehicle for US food aid shipments. Section 416(b) and PL480 shipments driven by surplus disposal objectives thus became the primary connection between food aid and civil conflict (Barrett and Maxwell, 2005; Schnepf, 2014). In this system, USDA wheat purchases (i.e., grain forfeitures for nonpayment of non-recourse commodity loans) were a function not only of the price, but also of the underlying loan rates. However, the only period during the NQ study window when loan rates fluctuated around market prices was a brief window between 1981 and 1986 (Westcott and Hoffman, 1999). This led government stocks of wheat to climb sharply, with government stocks reaching 62% of average annual wheat production 1981-1987 (Wescott and Hoffman, 1999). This represented a peak for government intervention in wheat markets, sparking changes to federal farm price support programs in the mid-1980s.

High levels of procurement during that period and the excessive stocks that resulted led to market reforms that de-linked wheat production and US food aid procurement, particularly following the Farm Bills in 1985 and 1990. Finally, the 1996 Farm Bill uncoupled the link between wheat production and government held stocks for good (Willis and O'Brien, 2015). CCC stocks of wheat were fully exhausted by 2006, and indeed, the Section 416(b) food aid program has been inactive since 2007 because of the unavailability of CCC-owned grain stocks (Schnepf, 2014). Since the early 1990s, the vast majority of US food aid has been procured by USDA on open market tenders announced in the Federal Register (Barrett and Maxwell 2005).

Effectively, this guaranteed the farmer the minimum of the rate fixed by CCC or the world price, and caused the CCC to purchase wheat when market prices were low. Government procurement was therefore a function not only of production and prices, but also of the level at which CCC set the loan repayment rate, a policy variable subject to revision in various Farm Bills.

Given that federal law began to unravel the link between wheat prices (and therefore wheat production) and government commodity procurement for use in food aid programs starting with the 1985 Food Bill and severed it in the 1996 Food Bill, if the mechanism NQ posit indeed drove their findings, then the first stage of NQ's IV strategy should be strongest prior to 1985 and non-existent after 1996. That turns out not to be the case. The post-1996 estimation offers a natural placebo test since the causal mechanism did not exist during that sub-sample.

Table C1 implements this simple robustness check by reproducing the first stage of the NQ strategy dividing the sample into three periods corresponding to the passage of the Farm Bills that successively decoupled US wheat production from government held wheat stocks. As expected, the connection between wheat production and food aid shipments is strongest prior to 1985 but statistically insignificant. In the years 1985 to 1996 the effect turns negative and statistically insignificant. The estimated relationship is inexplicably similar to the NQ baseline result in the post 1996 period but still not statistically significant. Indeed, there is no statistically significant difference between the sub-period when the pre-1985 sub-period when mechanism was in force and the post-1996 sub-period when it was not. The fact that we see a relationship between wheat

TABLE C1: RELATIONSHIP BETWEEN INSTRUMENT AND FOOD AID SHIPMENTS BY FARM BILL ERA

	(1)	(2)	(3)	(4)	(5)
VARIABLES	US food aid (1,000 MT)	US food aid (1,000 MT)	US food aid (1,000 MT)	US food aid (1,000 MT)	US food aid (1,000 MT)
Panel A: Pre-1985					
Baseline interaction instrument	0.00416	0.00261	0.00262	0.00216	0.00256
	(0.00290)	(0.00181)	(0.00195)	(0.00148)	(0.00175)
Observations	1,460	1,460	1,460	1,460	1,460
R-squared	0.73574	0.73809	0.75020	0.75813	0.76167
Panel B: 1985 to 1996					
Baseline interaction instrument	-0.00043	-0.00211	-0.00261	-0.00192	-0.00218
	(0.00132)	(0.00214)	(0.00274)	(0.00229)	(0.00246)
Observations	1,384	1,384	1,384	1,384	1,384
R-squared	0.60490	0.63565	0.65284	0.65952	0.66482
Panel C: Post-1996					
Baseline interaction instrument	0.00149	0.00039	0.00100	0.00165	0.00181
	(0.00083)	(0.00076)	(0.00094)	(0.00107)	(0.00114)
Observations	1,245	1,245	1,245	1,245	1,245
R-squared	0.69145	0.69852	0.72413	0.73353	0.73564
Controls (for all panels):					
Country FE	Yes	Yes	Yes	Yes	Yes
Region-year FE	Yes	Yes	Yes	Yes	Yes
US real per capita GDP x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
US Democratic president x. avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
Oil price x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
Monthly recipient temperature and precipitation	No	No	Yes	Yes	Yes
Monthly weather x avg. prob. Of any US food aid	No	No	Yes	Yes	Yes
Avg. US military aid x year FE	No	No	No	Yes	Yes
Avg. US economic aid x year FE	No	No	No	Yes	Yes
Avg. recipient cereal imports x year FE	No	No	No	No	Yes
Avg. recipient cereal production x year FE	No	No	No	No	Yes

Notes: The sample includes 125 non-OECD countries for the years 1971-1984 (panel A), 1985 to 1996 (Panel B), and 1997 to 2006 (Panel C). Coefficients are reported with standard errors clustered at the country level in parentheses.

production and wheat food aid shipments after the US formally ended the policy link that underpins NQ's identification strategy suggests that the first stage probably identifies off spurious correlation not related to the claimed exogenous policy mechanism. The fact that likewise calls into question the salience of the mechanism NQ use to support the relevance of the instrument, which could equally – and more plausibly? – arise from spurious time series correlation.

The NQ strategy does not hold only in the sub-period in which the true policy regime was closest to the one they describe, and not in the sub-period when that regime was not in effect. The failed placebo test enabled by exogenous policy change strongly signals that their results likely stem from some source of variation other than US commodity price stabilization and associated food aid policy. Furthermore, the type of policy regime they have in mind effectively ended in the mid-1990s, calling into question the current policy relevance of their results given dramatic changes in the way aid is distributed in recent decades.

Another type of placebo test might be feasible by exploiting variation among the US' various food aid programs. Food aid from the US is procured and distributed under multiple policies, each with its own legal authorization, priorities, and processes. Both historically and today, the bulk of food aid is distributed through PL480, which authorizes procurement and distribution of aid by USDA, along with distribution of Title II of PL480 by the US Agency for International Development (USAID) (Barrett and Maxwell, 2005). But PL480 consists of several titles which describe very different forms of aid.

Aid distributed by USDA through Title I of PL480 provides concessional sales of food aid directly to foreign governments. Recipient governments have historically sold off the vast majority of Title I food aid, treating these more as balance of payments transfers in kind than as food for direct distribution. Title I aid constituted the majority of food aid in the early period of NQ's study, accounting for 63% of US international Food Assistance Outlays between 1970 and 1979 (Schnepf, 2014). But the role of this direct-to-government concessional aid declined precipitously over the period and no allocations at all have occurred under Title I since 2006.

In contrast to Title I the role of aid distributed through Title II of PL480 has increased dramatically. Title II permits USAID to allocate aid in response to humanitarian emergencies and non-emergency food insecurity as an outright grant. Unlike Title I, the vast majority of Title II food aid shipments are directly distributed to food consumers. Today only the statutory minimum of 15% of Title II non-emergency shipments are sold ('monetized' in food aid jargon). Title II aid is delivered through non-governmental organizations (NGOs) and private voluntary organizations (PVOs) like CARE, Catholic Relief Services, or World Vision, or through intergovernmental organizations like the United Nations' World Food Programme (WFP) rather than directly to country governments. Title II accounted for less than 40% of US international food assistance outlays in 1970-1979, but accounts for 88% of food assistance today (Schnepf, 2014).

If food aid distributed to governments is more likely to fuel conflict (either because it is easier to steal or because governments use the food to feed their own troops or the proceeds of food aid sales to finance military operations) then it would might make sense to disaggregate food aid flows among programs and run the placebo test using Title II flows as a placebo. Because the transition from Title I to Title II food aid was largely coincident with the change in US farm policy that ended government wheat stocks – although these contemporaneous changes were driven by different political and economic phenomena (Barrett and Maxwell 2005) – and the data disaggregated between Title I, Title II, and other food aid program flows are not publically available, we defer just to the period-specific placebo test. One would expect that predominantly Title I PL480 food aid in the 1970s-1985 would be more likely to fuel conflict than mainly Title II PL 480 food aid in the post-1996 period. That would be consistent with the sub-period disaggregated effects we found in step 1 but that we could not corroborate in the sub-period placebo test just reported. But such a finding does not imply that the food aid distribution system that prevails today would have this effect; this would merely offer an explanation for a historical relationship.

Table C2 reproduces NQ's IV estimate of the relationship between food aid and conflict, but splitting the sample by the same periods as in Table C1. Consistent with the lack of a strong first stage relationship, we find no relationship between aid and conflict in

the period between the 1985 and 1996 Food Bills. As expected given that the instrument is strongest in the pre-1985 period, the coefficients on aid for this period are slightly bigger than those presented by NQ for the full sample. In the post 1996 period, the coefficient is also very close to the one in the NQ paper. The coefficient is not quite significant at standard levels, but given the smaller number of observations in that period, we cannot reject the null hypothesis of equivalence of the coefficients prior to 1985 and following 1996. Strikingly, despite the fact that food aid was administered very differently in the 2000s than in the 1970s, the coefficients for these two periods are nearly identical. Thus, one is left with two possible conclusions. Either, the delivery mechanisms behind food aid is irrelevant to the degree to which food aid translates into elevated conflict – which seems unlikely – or the NQ IV strategy is picking up something other than a causal effect of aid on conflict.

TABLE C2: 2SLS ESTIMATES OF FOOD AID ON CONFLICT BY FARM BILL ERA

VARIABLES	(1) Any Conflict	(2) Any Conflict	(3) Any Conflict	(4) Any Conflict	(5) Any Conflict
Panel A: Pre-1985					
U.S. wheat aid (tonnes) - from FAO	0.00218	0.00331	0.00327	0.00397	0.00357
	(0.00182)	(0.00294)	(0.00291)	(0.00322)	(0.00286)
Observations	1,460	1,460	1,460	1,460	1,460
R-squared	0.43882	0.20773	0.25901	0.12822	0.24196
KP F-Stat	2.050	2.078	1.813	2.140	2.138
Panel B: 1985 to 1996					
U.S. wheat aid (tonnes) - from FAO	0.00097	-0.00140	-0.00162	-0.00415	-0.00386
	(0.00744)	(0.00241)	(0.00206)	(0.00465)	(0.00408)
Observations	1,384	1,384	1,384	1,384	1,384
R-squared	0.60715	0.63300	0.63867	0.34016	0.41052
KP F-Stat	0.106	0.968	0.901	0.700	0.790
Panel C: Post-1996					
U.S. wheat aid (tonnes) - from FAO	0.00457	0.01265	0.00567	0.00368	0.00281
	(0.00301)	(0.02416)	(0.00600)	(0.00322)	(0.00258)
Observations	1,245	1,245	1,245	1,245	1,245
R-squared	0.56814	-0.43313	0.52204	0.65228	0.69256
KP F-Stat	3.219	0.256	1.141	2.384	2.527
Controls (for all panels):					
Country FE	Yes	Yes	Yes	Yes	Yes
Region-year FE	Yes	Yes	Yes	Yes	Yes
US real per capita GDP x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
US Democratic president x. avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
Oil price x avg. prob. of any US food aid	No	Yes	Yes	Yes	Yes
Monthly recipient temperature and Precipitation	No	No	Yes	Yes	Yes
Monthly weather x avg. prob. Of any US food aid	No	No	Yes	Yes	Yes
Avg. US military aid x year FE	No	No	No	Yes	Yes
Avg. US economic aid x year FE	No	No	No	Yes	Yes
Avg. recipient cereal imports x year FE	No	No	No	No	Yes
Avg. recipient cereal production x year FE	No	No	No	No	Yes

Notes: An observation is a country and a year. The sample includes 125 non-OECD countries for the years 1971-1984 (panel A), 1985 to 1996 (Panel B), and 1997 to 2006 (Panel C). Coefficients are reported with standard errors clustered at the country level in parentheses. In these shorter panels, collinearities arising when adding country characteristics cause fixed effects for several countries to be dropped, leading to the change in r-squared from column 4 to 5.

Step 4: Test for trend and difference stationarity

As appendix A explained, including year fixed effects does not eliminate the influence of seemingly spurious correlation between US wheat production and recipient country conflict because both variables display a similar inverted-U shape trend over the period, and, crucially, this trend is much more pronounced for regular aid recipients than for irregular ones. Including region-year and country fixed effects, even a time trend variable, does not permit causal identification unless either (i) the controls employed by NQ absorb all of the trend effects other than aid, or (ii) inter-annual US wheat production fluctuations and resulting food aid are in fact the dominant sources of the conflict trends.

These assumptions are much stronger than those described by NQ as the necessary and sufficient conditions for their strategy to reveal a causal effect of aid on conflict. Because the causal effect of aid on conflict is a relationship of substantial interest to policy makers, and because the clever econometric trick they employ appears in other empirical papers on other topics, we deem it important to highlight the caveats to the NQ conclusions and to demonstrate their vulnerability to spurious correlation due to nonlinear heterogeneous trends unrelated to the hypothesized mechanism.

NQ report results controlling for a linear trend, for lagged dependent variables, and using other conventional controls without testing for trend or difference stationarity in the underlying series. However, they never demonstrate that these are the correct specifications for the kinds of trends we should worry about. For example, controlling for a lagged dependent variable might be appropriate if the conflict series is non-stationary, but wheat production and wheat aid are stationary. Here we recommend the set of tests from time series literature that are appropriate for diagnosing issues of non-stationary processes and apply them to the NQ data.

When implementing a panel IV estimation where the variation comes from a time series, a reasonable sequence of steps authors might follow to assess the stationarity of the underlying data includes:

1. Conduct an augmented Dickey-Fuller (ADF) test on the uninteracted time series instrumental variable (i.e., lagged US wheat production). If the ADF test rejects

the null of a random walk or other non-stationary process, the authors should report the uninteracted IV specification including the F-statistic for weak identification, which should be at least 10. If the ADF test fails to reject the null hypothesis of a random walk, then spurious correlation is likely in the uninteracted IV regression case, and the authors should test for non-stationarity in panel variables.

2. Conduct a Hadri test on the interacted instrument, endogenous explanatory variable (i.e., wheat food aid), and outcome variable (i.e., conflict). If the test rejects the null hypothesis that all panels are stationary for either the interacted instrument or the outcome variable, then there is at least some risk of spurious correlation in the interacted IV specification.
3. Conduct a Fisher-type panel unit root test on the interacted instrument, endogenous explanatory variable (i.e., wheat food aid), and outcome variable (i.e., conflict). If the test fails to reject the null hypothesis that all panels have a unit root for one of these variables, this is strong evidence that this variable should be differenced until the subsequent test does reject the null appropriate to that order of integration. If the test rejects the null that all variables have the same unit root, one cannot rely on a single specification of a parametric time trend to solve the unit root problem that exists in only some of the country-specific time series. We are unaware of a test that can estimate the risk of mistaken inference from having a subset of countries follow a unit root due to a shared latent process. So we recommend assessing this risk through bootstrap and randomized placebo tests or simulation of the assumed data generating process.

We apply these steps to the NQ data by testing for non-stationarity in the lagged US wheat production variable. The workhorse method to test for non-stationarity in a single time series is the ADF test. ADF tests are based on regressions of the following form:

$$\Delta y_t = \alpha + \delta t + \Lambda y_{t-1} + \chi_t \quad \text{(C3)}$$

The null hypothesis is that y_t follows a random walk, which means that the lagged value y_{t-1} has no explanatory power about changes in y_{t-1} between periods t-1 and t. Failure to reject the null hypothesis that $\beta = 0$ therefore indicates failure to reject the random walk hypothesis. Different versions of the ADF test impose restrictions that $\alpha = 0,\ \delta = 0$ (no constant), $\alpha \neq 0,\ \delta = 0$ (constant), or $\alpha = 0,\ \delta \neq 0$ (trend). An additional version imposes a modified null hypothesis that y follows a random walk plus non-zero drift, which

uses the same regression model as the constant form of the test but implies different critical values.

Table C3 below shows the critical values for each version of this test using the data originally included in NQ. Because conflict and wheat food aid are different for each country, we average these variables across countries at the yearly level into a single time series. The test is therefore whether the proportion of countries experiencing conflict or the average wheat food aid receipts follows a random walk. Only the drift model rejects the random walk plus non-zero drift for all versions of the test. When we impose the hypothesis that these variables follow a random walk without drift, we cannot reject the null for conflict and wheat food aid for any version of the ADF test.

Table C3: ADF Test Results on NQ Data

	No Constant	Constant	Trend	Drift
US Wheat Production in year t-1 (metric tons)	-0.088	-3.543**	-3.376*	-3.543***
Proportion of countries experiencing conflict in year t	-0.297	-1.660	-1.503	-1.660*
Average receipt of US wheat food aid in year t (metric tons)	-1.271	-2.014	-2.443	-2.014**
1% critical value	-2.644	-3.682	-4.288	-2.445
5% critical value	-1.95	-2.972	-3.56	-1.692
10% critical value	-1.604	-2.618	-3.216	-1.308

Notes: Reports ADF test coefficient estimates for the three variables indicated in rows. *,**, and *** indicate rejection of the unit root null hypothesis at the 10, 5, and 1 percent significance levels, respectively, relative to the appropriate critical values indicated in the bottom rows. Columns indicate the ADF model specified for the critical value. All tests run with Stata's dfuller command. N=35 for each test, covering the years in the 1971-2006 time series NQ use.

Since we cannot reject the null hypothesis that lagged US wheat production follows a random walk without drift, spurious correlation becomes a concern. Any trend among regular aid recipients that is not also common to infrequent recipients of aid could be spuriously correlated with the dynamics of wheat production. Even if controlling for the trend or lagged dependent variables corrects this non-stationarity, the F-statistic for weak

instruments in the uninteracted case is well below 10 (F=1.83 to 3.44 over reported specifications), meaning that the uninteracted specification is vulnerable to both spurious regression and weak instruments problems.

Interacting lagged US wheat production with the country-specific regularity of food aid receipt transforms the instrument from a time series variable into a panel variable. Conflict and aid receipts are both also panel variables as they vary both within countries over time and within years across countries. The question then becomes whether a spurious regression problem exists in the panels and whether existing time series tests can be used to diagnose these problems.

A range of tests exist to diagnose unit root processes in panel data. The ideal test would tell us the following four features of the data for each of the three main variables:

1. Do any countries show evidence a unit root or other non-stationary dynamic process in their time series for this variable?
2. What share of countries show evidence of a non-stationary process?
3. What is the expected rate of over-rejection using conventional inference tests (t-tests for significance of first stage and reduced form, F test for weak identification) if not accounting for the share of countries whose time series may have nonstationary processes?
4. If dynamics are correlated across countries, what is the size of the possible bias that could result?

Unfortunately, existing tests generally are based on sharp null hypotheses. A Hadri test statistic relates to the null hypothesis that the time series for *every* (country-specific) panel is stationary, providing a test for question (1). Applied to the NQ data, this test rejects the null hypothesis for all three main variables, conflict, quantity of wheat food aid received, and lagged production of US wheat interacted with regularity of food aid receipts, each with $p<0.0001$. This tells us that were we to implement the IV strategy by regressing conflict and wheat food aid on lagged US wheat production for one country at a time, at least one such regression would suffer from spurious correlation. Unfortunately, this test does not reveal the share of countries whose time series reflect a unit root nor does it convey information about the extent of any inference and bias issues which could result, leaving questions (2) and (3) open.

By contrast, a Fisher-style panel unit root test computes the ADF statistic for the time series of each panel one at a time and then provides a test statistic for the null hypothesis that a unit root process exists for *all* time series. The Fisher-style test rejects the null of a unit root for *all* panels. Given that the Hadri test gives strong evidence that at least one country has a unit root process for each of the main variables, the Fisher-test tells us that different countries follow different processes. They neither are all stationary nor all following the same unit root process. Failure to reject the null with a Fisher-style test would provide strong evidence that each of the key variables should be differenced until stationary. In the NQ data, the Fisher-style test with one lag rejects the null for each of the major variables: conflict, wheat food aid deliveries, and lagged US wheat production interacted with regularity of aid receipts, each with $p<0.0001$.

While this rejection does not provide strong guidance in favor of differencing the variables, it also means that controlling for common time trends through either parametric specifications of time effects or non-parametric year "fixed effects" are unlikely to address the risk of spurious correlation. Because not all variables have the unit root, the unit root processes cannot be common to all countries and thus has no clean corrective.

In summary, the guidance from existing panel tests for unit roots is ambiguous in the NQ data. The ADF test on wheat food aid suggests that uninteracted US wheat production should not be used alone as an instrument because of the risk of spurious correlation. The Hadri test suggests that NQ's choice to interact with frequency of aid receipts does not fully address the problem, because we cannot reject a unit root for at least one country. The Fisher-style test tells us that controlling for common time trends alone will not address problems that arise due to unit roots in some, but not all, of the country-specific time series. Because not all countries share the same unit root, solutions that assume all countries share the same unit root – such as such as fixed effects –will not solve the problem.

Unfortunately, all existing tests we are aware of measure neither the share of countries whose time series follows a unit root or other nonstationary process nor the risk to inference or bias that would arise from a given share. We can conduct the ADF test

separately for each main variable and each country and report the share where we fail to reject the null of random walk for that country. For the main variables, using an ADF test with no drift, the share of countries in which we fail to reject the null of a random walk at a 5% level are 57% for conflict, 53% for wheat food aid, and 12% for the interacted instrument. This offers no support for the absence of correction for nonstationarity in the original NQ estimates.

Table C4: ADF test results by regularity of food aid receipt

	Below median food aid receipt regularity	Above median food aid receipt regularity
Average proportion of years in which aid was received by countries	0.10	0.62
Average proportion of years in which countries experience a war	0.15	0.27
Average quantity of wheat food aid received (MT)	1.67	49.27
Average value of interacted instrument (regularity of aid receipts x lagged US wheat production (MT))	5940.47	36508.89
Proportion of countries whose ADF test on the **conflict** time series has p >.05	0.54	0.59
Proportion of countries whose ADF test on the **interacted instrument** time series has p >.05	0.11	0.12
Proportion of countries whose ADF test on the **wheat aid** time series has p >.05	0.61	0.45

The last natural question is whether these tests suggest that the unit root process is equally common among regular recipients and irregular recipients. If these patterns were similar, we might think that any bias or inference issues would "difference out" by comparing regular versus irregular recipients. Table C4 shows the share of countries where we fail to reject the null hypothesis of a random walk for each variable, splitting by whether regularity is above or below median. The rates of rejecting the random walk hypothesis in these data do not differ dramatically between above and below median regularity of aid receipts. However, such a comparison does not tell us what difference would be big enough to cause a problem or whether the median is the right stratification point for comparison.

Because existing panel IV tests impose sharp null hypotheses, they can only signal a risk of spurious correlation, but do they cannot estimate the risk of bias or mistaken inference due to spurious regressions. We therefore turn to methods based on bootstrap, placebo randomization, and Monte Carlo simulation to study the risks to inference and bias in these data.

Given that time series tests reveal a risk of nonstationarity in the primary variables, an assessment of the NQ results should report which trends can be distinguished from NQs reported results. As we demonstrated in Table 3, re-estimating the NQ model using first-differenced series overturns their core findings. But even if one just includes a quadratic rather than a linear trend in their regressions, without first-differencing, their results evaporate. The confidence intervals explode to include both coefficients that are more than 10 and -10 times their coefficient estimate, and a weak instruments test signals a high risk of spurious correlations (Cragg-Donald F statistic = 0.242). This result implies that any effect NQ identify cannot be distinguished from any other that might exhibit a quadratic trend, including both random cycles arising from summed errors over years or omitted variables with quadratic trends.

Furthermore, in the specifications NQ report for the uninteracted case, all reported F-statistics are below four, suggesting a weak instrument. When the interaction is introduced, the reported F-statistic increases above the usual benchmark of ten for two out of five reported specifications with different controls. In our simulations, the interaction does not eliminate the bias from spurious correlation due to trends. Searching for specifications with different interactions that could plausibly be linked to the proposed instruments creates a risk that instruments that are spuriously associated with the causal variable are reported with an artificially high level of confidence. Another way to put this is that interactions give researchers degrees of freedom to search for "splits" of the data that pick up those panels at risk of spurious correlation to inflate the F-statistics. To give a simple example in the NQ data, over 31.5% of countries do not experience any conflict in the dataset and 14.2% of countries never receive any wheat aid. In a model without year fixed effects, the time series for these variables will be strongly stationary,

and so not will not be spuriously correlated with wheat production. Interacting wheat production with a variable that downweights countries with no aid or conflict will, however, increase the F-statistic by increasing the weight placed on the panels where spurious correlation is a real risk.

Step 5: Run Young's bootstrap test for overleveraged observations

As explained in the main text, the appropriate concern Young (2018) raises about the possibility of over-leveraged observations merits exploration in any IV estimation study. But that problem is quite distinct from the spurious regressions problem that motivates this paper.

NQ's results pass Young's test. Here we report what Young calls the bootstrap-c, which equals one minus the share of bootstrap iterations for which the coefficient estimated in a given bootstrap sample is greater than zero. Figure C7 shows the distribution of 2SLS estimates obtained by dropping each of the 127 countries one at a time and estimating the NQ 2SLS without those observations. This procedure always returns a positive coefficient of aid on conflict when excluding any single country from the analysis. The bootstrap-c p-value when bootstrapping over countries is 0.048, which is greater than NQ's reported p-value, but still significant at the five percent level. We favor running leave-one-out tests at the country level because that's the unit NQ cluster. If we instead treat the relevant clusters as years rather than countries, similar results obtain. The coefficient estimates are never negative when leaving any one year out, and the bootstrap-c p-value is 0.018. So leveraged country observations is not an issue in the NQ 2SLS estimates.

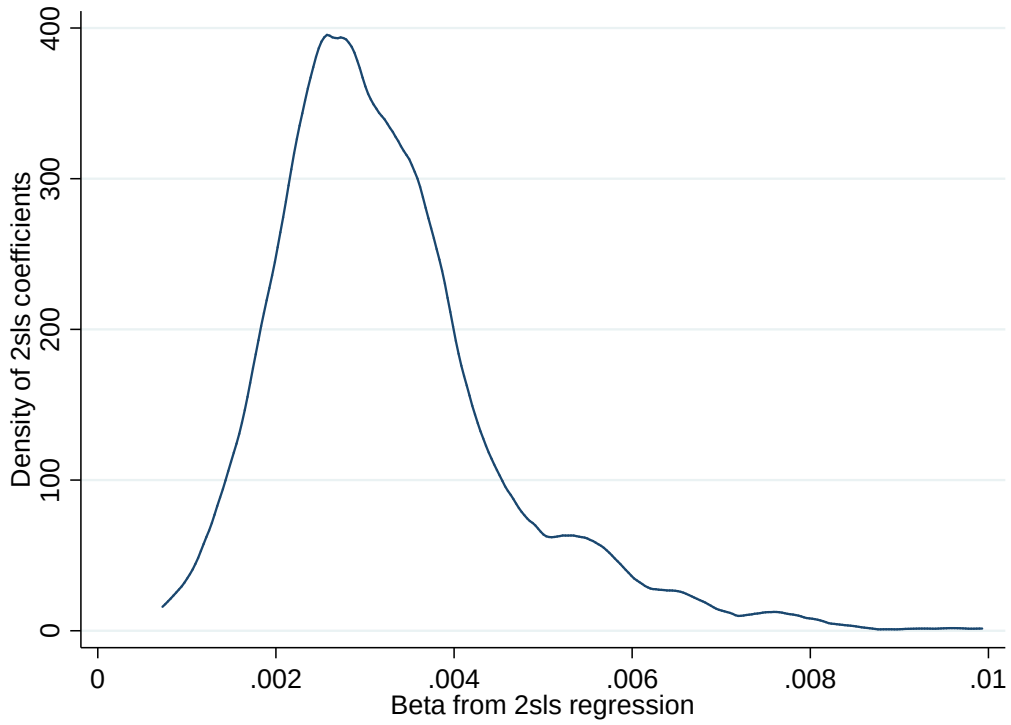


Figure C7: Kernel density estimate of Young's bootstrap-c p-values

Step 6: Check the instrument using a randomized placebo test

The randomization placebo test we introduce rests on the simple principle that introducing randomness into the endogenous explanatory variable of interest (in the NQ case, a country's food aid receipts in a given year) while holding constant the (potentially endogenous) cross-sectional exposure variable ($\overline{D}_{ir}$), the instrument (US wheat production) and everything else should eliminate the estimated causal relationship if indeed exogenous inter-annual shocks to the endogenous explanatory variable (wheat food aid shipments) drive outcomes (conflict in recipient countries). So within a given year, we hold constant the following variables: the quantity of wheat produced, the identity of the countries that receive any wheat food aid from the US (thereby fixing both $\overline{D}_{ir}$ and the timing of food aid receipts), observable fixed and time-varying characteristics of countries, and the aggregate distribution of wheat food aid allocations across all countries each year. But we randomly assign the key variable of interest, the quantity of

aid delivered to *a particular country*. For example, in 1971, 60 countries received any wheat food aid from the US. In our simulation, we randomly reassign (without replacement) the quantity of wheat aid deliveries among these 60 countries, while holding constant the (true) zero value of food aid receipts in the other countries. For example, instead of receiving the 2,100 tons it actually received in 1971, Nepal could be randomly assigned the 800 tons actually shipped to Swaziland that year. We similarly reshuffle the wheat aid allocations among the 62 countries who received aid in 1972, and so on for every year in the sample.

The resulting pseudo-data set preserves the two sources of endogeneity we worry about – time trends and endogenous selection into being a regular food aid recipient – but sweeps out the source of variation that NQ have in mind by randomizing among countries the assignment of specific food aid shipment volumes. To keep with the earlier example, Swaziland's food aid receipts cannot plausibly have caused civil conflict within Nepal. This way, conflict can remain spuriously related to wheat production because neither the conflict time series nor the wheat production time series nor the exposure variable that distinguishes between groups are altered, but the causal mechanism has been rendered non-operational by randomization since it is no longer the case that in expectation particular countries receive the randomly generated additional aid in a given year. In this placebo test, the only reason why the quantity of wheat aid delivered would be positively related to conflict in NQ's baseline 2SLS specification would be that countries that regularly experience conflict are also the countries that regularly receive food aid (which is what we would expect if aid were targeted to humanitarian crises) and the years of high wheat production happen to be years in which conflict is elevated (which with only 36 years and strong trends could well be spurious).

Figure C8 shows the distribution of coefficient estimates generated by 1,000 randomizations of food aid allocations and then (re-)estimating the baseline 2SLS model. If the true causal relationship between food aid allocations and conflict were positive and the identification was otherwise unaffected by selection bias and spurious time trends, the distribution of coefficients would shift left relative to the NQ coefficient estimate – and if

the share of countries in which aid causes conflict is small relative to a large enough sample, would center around zero – because the randomization of food aid allocations would attenuate the estimated relationship between aid and conflict. Instead, we find the opposite. The distribution of parameter estimates clearly shifts to the right of the NQ 2SLS coefficient estimate. This implies that the identity of aid recipient countries and the overall trends in global conflict prevalence, US wheat production, and total food aid deliveries drive the estimated relationship, not inter-annual fluctuations in food aid receipts by a given country. Indeed, to the extent that the IV does contain some component of random aid allocation, this test also signals that the true association between inter-annual variation in food aid receipts and conflict must be negative since eliminating that source of variation causes an increase in coefficient estimates.

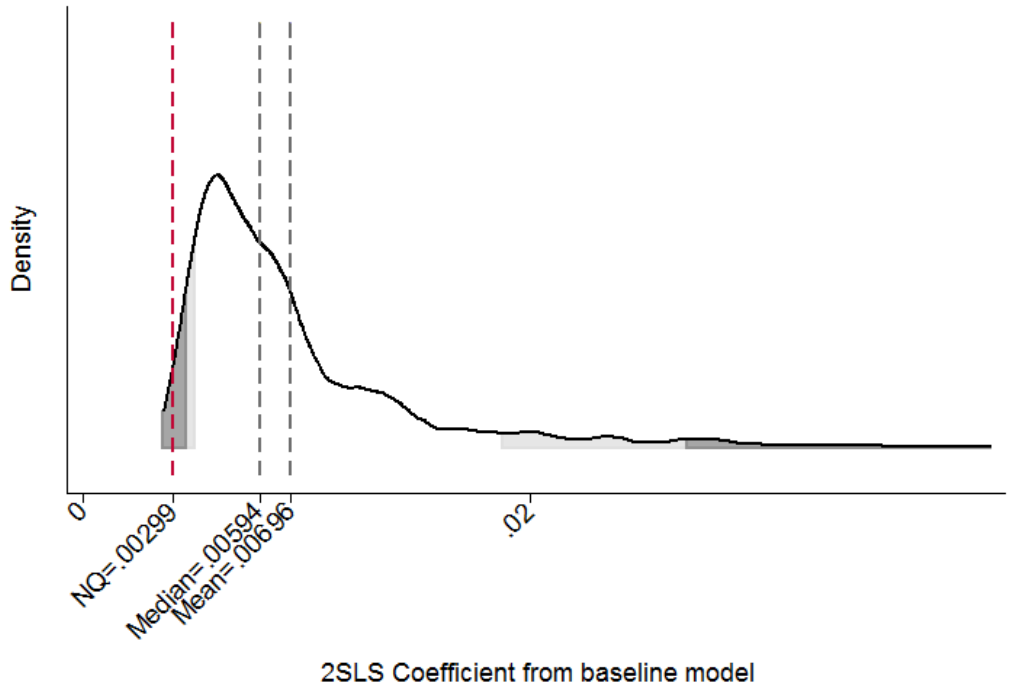


FIGURE C8: DISTRIBUTION OF 2SLS COEFFICIENT ESTIMATES USING RANDOMIZED FOOD AID ALLOCATIONS

Notes: The density plot depicts the distribution of 2SLS coefficient estimates using the set of baseline controls with 1,000 draws of randomized allocations of wheat aid food volumes among actual recipients in a particular year. The dark shaded area indicates the bottom and top 5% of draws. The light shaded area shows the top and bottom 10%. The kernel density function and percentiles are estimated on the full set of 1,000 iterations, but the plot trims the largest 20 and smallest 15 values of the distribution for scale.

Step 7: Monte Carlo placebo tests: simulation after eliminating or reversing the causal mechanism

As the placebo tests based on randomizing food aid allocations demonstrate, what really seems to drive the NQ findings is the fact that conflict incidence follows nonlinear trends that differ between irregular and regular recipients of aid, with the pattern for regular recipients co-cyclical with US wheat production and food aid shipments. As we show in section I of the main paper, any variable that exhibited a similar inverted-U (or U)-shaped time series pattern over the sample period would "work" as an instrument, no matter how spurious the association with food aid shipments. In order to show that NQ's IV approach identifies only spurious correlation and not a causal effect, we now build a simple model that replicates key features of the underlying data, but intentionally break – and then reverse – the hypothesized causal relation between food aid flows and conflict. This allows us to know the true data generating process (DGP). We then generate pseudo-data from that DGP and replicate the NQ estimation strategy 100 times to generate a distribution of estimation results. When the DGP by construction has food aid statistically independent of conflict, this provides an attractive placebo test.

Suppose that wheat production follows the following pattern:

$$Wheat_{it} = f(t) + h_t \quad \text{(C4)}$$

Where f(t) is a function of time and h is a random variable distributed N(0,1) that is independently distributed across years. Define a random variable f_t =f(t) for a fixed time period that over the study period where the variance of f_t over the studied period, $\sigma_{f,}$ is substantially larger than the variance of h_t, σ_h. In such a framework, observations of wheat production that are temporally proximate are strongly related due to the underlying trend, f_t, while most of the idiosyncratic variation in wheat production occurs as relatively moderate deviations around the mean. This describes a basic pattern of the NQ data, as shown in Figure 3 in the main paper. In our simple model conflict occurs whenever a latent

variable representing conflict risk exceeds a threshold, where conflict risk is also subject to trends as follows:

$$Conflict_{it} = 0 \text{ if } R_{it} <$$

(C5)

$$Conflict_{it} = 1 \text{ if } R_{it} > \bar{R} \text{ where } R_{it} = a_i * v_t$$

(C6)

and a_i is a random variable uniformly distributed between 0 and 1 and specific to countries which captures each country's vulnerability to outbreak of conflict induced by a globally common shock $v_{t.}$. We refer to a_i as fragility as it represents a country's specific risk to factors that affect conflict. v_t is the globally common shock to countries and consists of a trending component and a temporally idiosyncratic component such that $v_t = g(t) + j_t$. As with wheat production, $j_t \sim N(0,1)$, and we consider a context where if we define g_t as a random variable defined by g(t) for a range of years, then within the study period σ_g is substantially larger than σ_j. Modeling conflict this way matches a feature of the data, as described in step 4 above, that conflict prevalence is similar in adjacent periods and a large portion of the time series variance is explained by long-term trends.

The main worry in identifying the causal effect of aid on conflict is that aid may be directed toward countries that experience additional conflict. To capture this concern, we model aid, following stated US government policy, as determined by conflict and assess whether NQ's IV strategy removes the bias from this source. To capture the simultaneous relationship:

$$aid_{it} = Max(0, Conflict_{it} * U_{it})$$

(C7)

where $U_{it} \sim N(0,1)$. The function describing aid allocation has three features. 1. Countries that are not experiencing conflict never receive aid, i.e., aid is only sent to countries

experiencing conflict,[30] 2. Some countries that experience conflict do not receive aid for exogenous reasons on account of a low draw for j_{it} (which can be thought of as features of the current political relationship with the US, for example). 3. The amount of aid received conditional on non-zero aid receipt can be random and independent of a country's risk of conflict and can be determined by factors other than conflict, such as politics or ease of transport, here modeled as a high draw for j_{it}.

To show how the time trends arising from the functions f(t) and g(t) can influence the NQ results, we choose functional forms for these components that correspond approximately to patterns observed in the data, and show through Monte Carlo simulations that NQ's results are reproducible without the need to assume any causal effect of aid on conflict or any correlation beyond those above. Indeed, below we even introduce a negative causal effect of aid on conflict – the opposite of what NQ claim to find – and show that we can still replicate their findings.

In the first simulations, we assume that $g(t) = f(t) = t-(1/36)*t^2$ where t=1… 36. This form forces both prevalence of conflict and wheat production to follow an inverted "U" shape over a time horizon of 36 periods, the number of years in the dataset. The plots below simulate one draw of a pseudo-dataset with 126 countries and 36 years for this model to show how the conflict and wheat production dynamics work. Figure C9 shows how the secular trend g(t) and the country-specific conflict vulnerability combine to determine conflict risk. Dark shaded dots are countries with high fragility, a_i, while lighter shaded dots are less fragile countries, those with a low a_i. Within any year, some countries are at greater risk of conflict than others because of their higher fragility, represented by the fact that dark dots always appear higher than light dots on the vertical axis. Any dots above the horizontal line are those whose conflict risk exceeds the threshold and are considered to be in active conflict. Across years, the overall risk of conflict is increasing in early periods and falling in later periods as shown by the shift of the dark dots up the vertical (conflict

[30] This is an oversimplification of the real aid process. In the NQ dataset, countries that are not experiencing conflict also receive aid, but the percentage of countries receiving aid is substantially and statistically significantly higher in countries that are experiencing conflict than in countries that are not (47.0% vs. 34.6%). We will soon relax this assumption to show that allowing components of aid to be unrelated to conflict does not change the estimation results and can explain other features of the results reported by NQ.

risk) axis in the first half of the period and shift down in the second half. The consequence of this DGP is that some (high fragility) countries nearly always experience conflict and some low fragility ones are almost never in conflict, but because of the secular trend, the countries with similar conflict risk tend to enter and exit the conflict state in the same years. Once a country enters conflict, it tends to remain so until global conditions improve enough to bring it back below the conflict threshold.

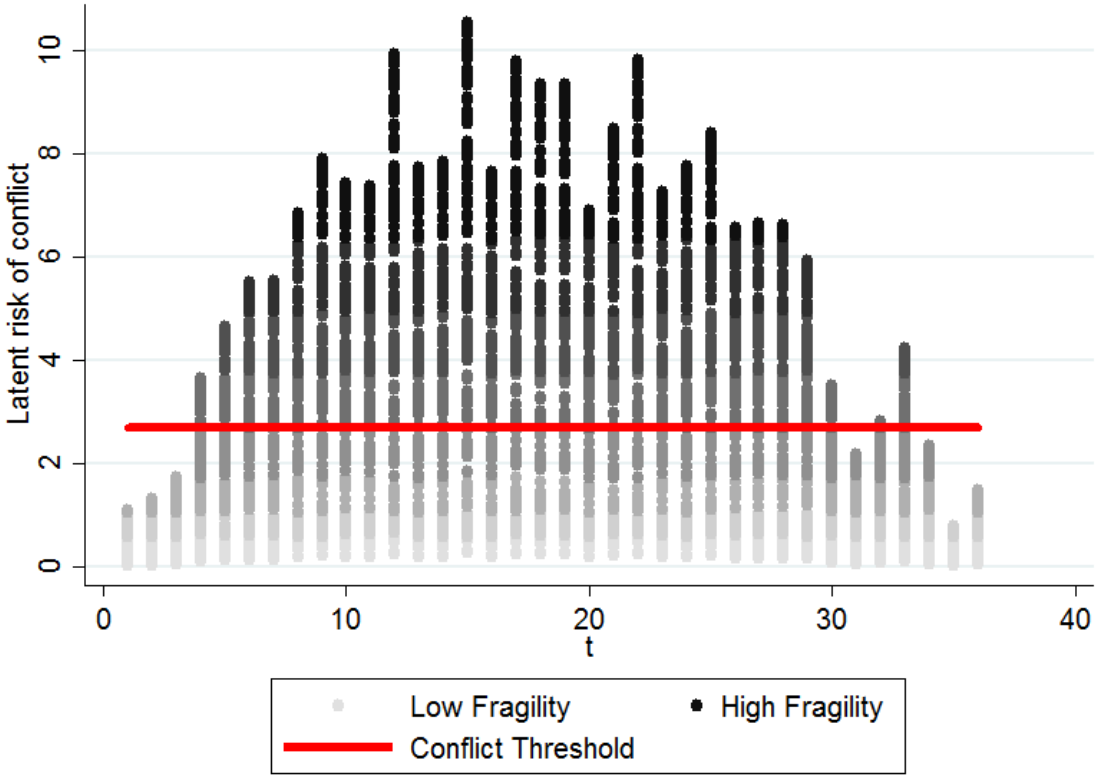


FIGURE C9: SIMULATED RISK OF CONFLICT OVER TIME BY FRAGILITY LEVEL IN THE MODEL

Because we imposed an inverted "U" shaped trend for g(t) and f(t), conflict and wheat both follow a pattern of starting and finishing the period at low levels and peak in the middle of the period, as shown in Figure C10. This corresponds with observable patterns in wheat production and conflict in the 36-year period in the NQ data. Because we have chosen forms such that $\sigma_g > \sigma_j$ and $\sigma_f > \sigma_h$, the variance in both conflict and wheat is

dominated by the trend rather than by year-on-year idiosyncratic variation around the trend, as is true in the actual data.

In this model, there is no true causal effect of aid on conflict in the DGP. If there is a correlation between aid and conflict, it comes from aid agencies preferring to send food aid to more conflict prone places. We now replicate NQ's identification strategy by drawing 100 random

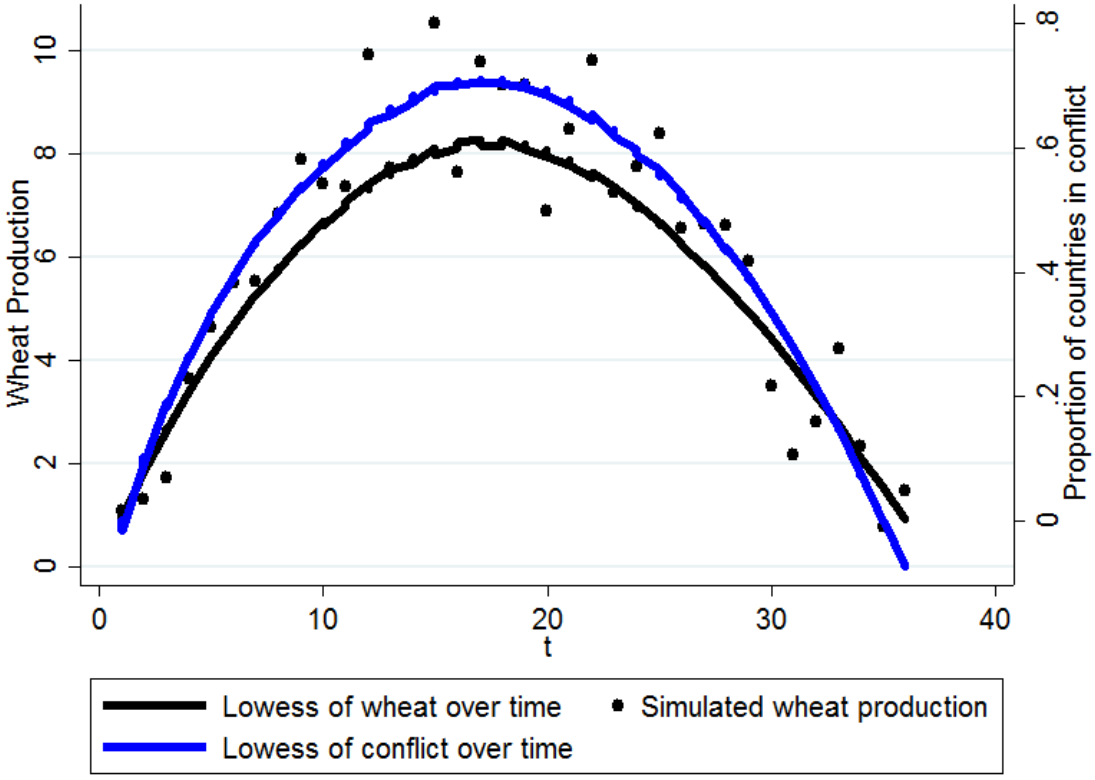


FIGURE C10: SIMULATED WHEAT PRODUCTION AND AVERAGE CONFLICT RISK IN THE MODEL

samples of 36 years and 126 countries using the model above and then estimate the relationship using 2SLS that mirrors NQ, with the following first state and reduced form equations:

$$aid_{it} = \pi^{mc} wheat_{it-1} * d + C_i + Year_t + M_{it} + \eta_{it}^{mc}$$

(C8)

$$conflict_{it} = \gamma^{mc} wheat_{it-1} + c_i + year_t + m_{it} + \mu_{it}^{mc}$$

(C9)

where C_i and c_i are country fixed effects, $Year_t$ and $year_t$ are year fixed effects, and M_{it} and m_{it} are country and year level controls. Figure C11 plots the density of estimated β coefficients from the second stage for two specifications. The "parsimonious controls" specification includes only time and country fixed effects and the "lagged dependent variable" specification includes conflict in period t-1 as an additional control. This corresponds to the estimated coefficient of aid on conflict in the parsimonious specification of NQ's IV strategy. The distribution of estimated instrumental variables estimators $\frac{\widehat{\gamma^{mc}}}{\widehat{\pi^{mc}}}$ never includes the true zero value; rather all the estimates of the relationship between aid and conflict are positive. The estimator clearly fails the placebo test. Including the year and time (region-year in NQ) fixed effects or the control for lagged conflict does not prevent the IV coefficient estimate from picking the up the bias that arises from the endogenous determination of aid and conflict and the spurious correlation of the time series.

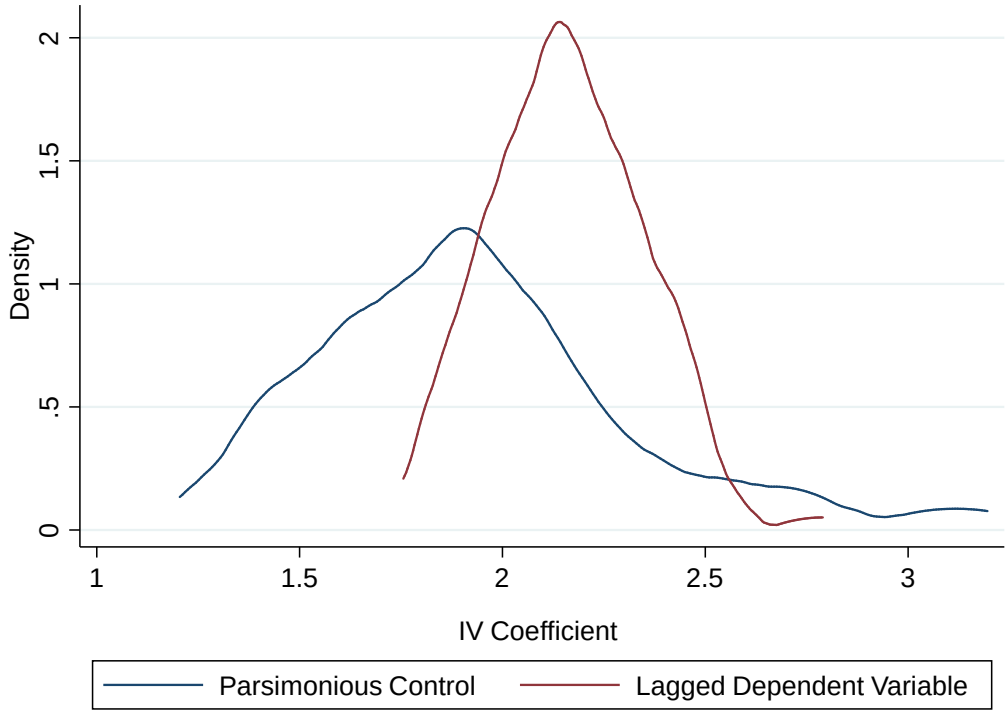


FIGURE C11: DISTRIBUTION OF SIMULATED COEFFICIENTS FOR IV SPECIFICATION OF SIMPLE MODEL

Notes: The plot displays the distribution of 100 simulated 2SLS coefficient estimates on instrumented aid from the IV specification. The parsimonious specification includes year and country fixed effects in both the first and second stage. The lagged dependent variable specification adds that AR(1) term to the parsimonious specification.

This simple model suffices to generate a strictly positive estimated relationship between aid and conflict even in the absence of a causal effect of aid on conflict. But a few additions are necessary to describe other key features of the data. In particular, in the simple model, the OLS relationship would always be positive, because in this model aid never flows to places that never experience conflict.[31] But in the actual data, NQ find a negative, albeit statistically insignificant, OLS estimate associating food aid flows with conflict. NQ argue that the positive IV relationship is generated by a causal effect of aid on conflict while a small negative and insignificant OLS coefficient estimate is explained either by

[31] In our simulations of this baseline model where aid and conflict have a joint inverse-U trend, the IV estimates are always upward biased relative to OLS for the reasons explained in the main paper.

measurement error[32] or following an unsubstantiated hypothesis that food aid programs are on average effective at targeting aid away from places where conflict is prevalent or at least away from places where aid causes conflict. Of course, the idea that USAID directs aid away from places at risk of conflict directly contradicts the stated objective of the agency and the patterns in the data, making this an unattractive candidate explanation.

Simple adjustments to our simulation model can generate the observed negative OLS coefficient and an upwardly biased IV estimate without the need to assume a causal effect of aid on conflict. In fact, the mechanism that generates the relationships in the extended case is an assumption that aid *prevents* some conflict, but coincident trends generate spurious correlation that dominates the IV estimation and obscures the true negative causal relationship between food aid flows and conflict. If one keeps all the components of the model but allows for some component of aid to be unrelated to conflict and allows for food aid flows to reduce the risk of conflict, the negative OLS relationship found by NQ can be generated by the model as well.

To show this, we modify the aid and conflict dynamics as follows:

$$Aid_{it} = Max(0, I(R_{it} > R)*j_{it}) + s_{it} \quad \text{(C10)}$$

$$\text{Where } s_{it} = \begin{cases} 2*Max(0, n_{it}) \text{ with probability .5} \\ 0 \text{ with probability .5} \end{cases}$$

$$n_{it} \sim N(0,1)$$

The new aid function allows the possibility that not all food aid is determined by conflict. Rather, food aid flows can vary on both the intensive and extensive margin in idiosyncratic ways, for example because aid may be shipped as a response to natural disasters. The component of aid that arises for these reasons is described by s_{it}, which has two sources of idiosyncratic variation. The first partrepresents the part that is correlated

[32] In order for measurement error in the conflict variable to drive the results, the measurement error would have to be systematic in such a way as to actually flip the sign of the observed relationships. It is not clear what would generate such measurement error in this context.

with the quantity of the food aid shipment, for example, the intensity of a natural disaster. The second part is the exogenous reasons that a country with a positive draw of n_{it} will receive any aid or not. This is modeled by the fact that a given country in a given year that has a positive draw of n_{it} has a 50% chance of receiving aid and a 50% chance of not receiving aid.

In contrast to the earlier model, there are now two reasons that countries receive food aid. The proportion of countries experiencing conflict will vary by year, but in expectation, 50% of the countries experiencing conflict will receive some aid because of their conflict status. Aid for non-conflict reasons such as disasters does not systematically vary by time. In expectation, 50% of countries will experience a disaster that warrants aid shipments, and 50% of the countries that experience such a shock receive aid.

The second modification is that aid flows can prevent conflict, as food aid advocates hope. We now modify the model so that if a country receives aid for any reason, either because of a high risk of conflict or because of a disaster, it will not experience conflict in that year. Thus, all countries that would experience conflict (those with $R_{it} > R$) can receive aid, but only some receive it. Any countries that receive aid for this reason or because of the idiosyncratic component w_t will not experience conflict in that year so that:

$$Conflict_{it} = 0 \text{ if } Aid_{it}>0 \quad \text{(C11)}$$

These modifications to the model capture several plausible features of the true DGP:

1. Countries that are at risk of conflict are likely to get aid for that reason.
2. Countries also receive aid for idiosyncratic reasons that are not related to conflict, such as political considerations, natural disasters, etc.
3. Aid can prevent conflict, for example by alleviating the conditions that cause people to fight or because delivery of US food aid is associated with actions

and investments that control conflict, such as the deployment of peacekeeping troops.

Modifying the model in this way and repeating the Monte Carlo simulation 100 times leads to the distribution of OLS and IV coefficients shown in Figure C12. The OLS estimates from NQ's specification are now negative in all simulations, just as NQ find, reflecting the fact that although countries that experience conflict receive the most aid, they experience conflict only in the years when for exogenous reasons they do not actually receive aid. The negative OLS estimates capture the effects of aid flows that arise both because of conflict risk and other unrelated sources. The IV relationship remains positive due to the spurious regressions problem, just as in NQ. The instrument misleadingly focuses on differences in aid that are endogenous to conflict for two reasons. Countries that experience the most conflict are most likely to get aid and conflict and wheat production are spuriously correlated over time.

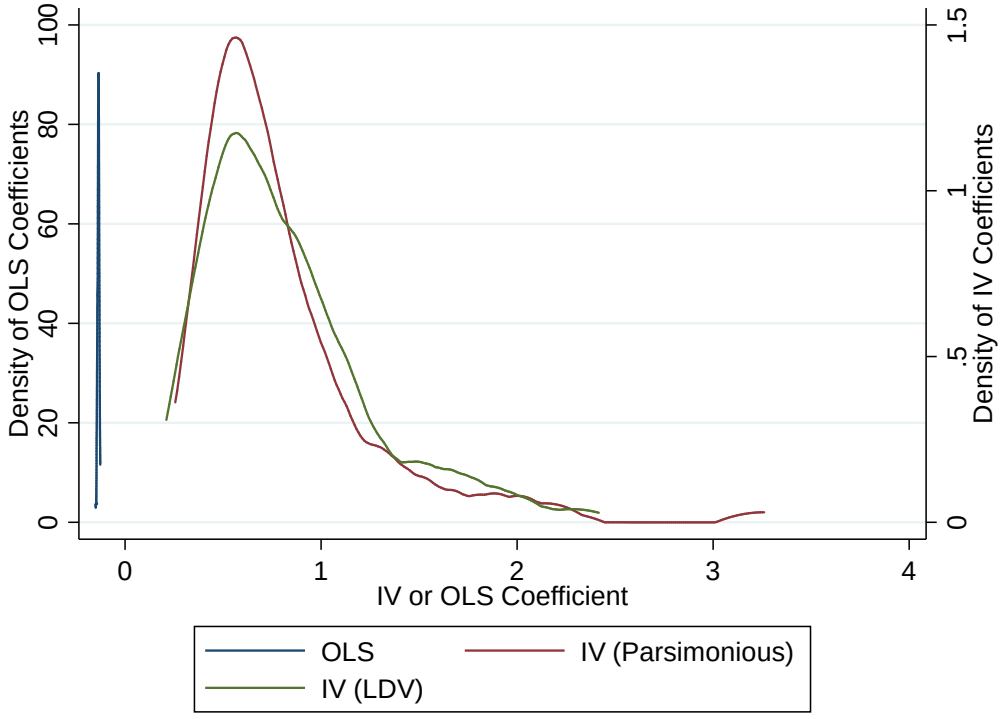


FIGURE C12: DISTRIBUTION OF SIMULATED COEFFICIENT ESTIMATES FOR EXPANDED MODEL

Notes: The figure displays the distribution of 100 simulations of the OLS coefficient estimates along with 2SLS estimates under two specifications. The parsimonious specification includes year and country fixed effects in both the first and second stage. The lagged dependent variable specification adds that AR(1) term to the parsimonious specification. To keep the scale readable, the figure is produced dropping one coefficient estimate <-9 on the IV with lagged dependent variable .

These results highlight the fact that NQ's comparison of OLS and IV results does not compel a conclusion that food aid causes conflict. In fact, in the expanded model above, the same parameter estimate patterns they find appear in a context where aid actually *prevents* rather than causes conflict. So not only do the NQ estimates fail the Monte Carlo placebo test, they fail to replicate a known causal mechanism that reverses the direction of causality from the one they hypothesize. This directly reflects the problem, explained in the main paper, that spurious regression in the presence of reverse causality will bias panel IV estimates in the direction of the reverse causal correlation between the outcome variable (conflict) and the endogenous regressor aid), in this case, positive bias to the 2SLS estimates.

This Monte Carlo exercise underscores that one needs to carefully explore the longer-run time series patterns that might overwhelm the inter-annual variation used to identify true causal effects, thereby generating spurious correlation in panel IV estimates such as NQ's. Because wheat production and conflict in regular aid recipients both show pronounced, parallel trends in the time series, but the conflict incidence trend in irregular aid recipients does not parallel that of regular aid recipients – and aid receipt is endogenous – NQ's estimation strategy falls prey to the spurious regressions problem outlined in the main paper. This Monte Carlo result reinforces the conclusion of the preceding placebo tests.

APPENDIX REFERENCES

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 Kevin C. Heisey (2002). "How Effectively Does Multilateral Food Aid Respond to Fluctuating Needs?" *Food Policy.* 27(5-6): 477-491.

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

Blattman, Christopher and Edward Miguel (2010). "Civil War." *Journal of Economic Literature.* 41(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.

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

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2018), "Bartik Instruments: What, When, Why, and How", NBER working paper 24408.

Jayne, Thomas S., John Strauss, Takashi Yamano, Daniel Molla (2002). "Targeting of Food Aid in Rural Ethiopia: Chronic Need or Inertia." *Journal of Development Economics,* 68: 247-288.

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

Schnepf, Randy (2014). "International Food Aid Programs: Background and Issues." Congressional Research Service.

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

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

USAID. "Office of Food for Peace." 5 March 2015. Accessed 15 May 2015. http://www.usaid.gov/who-we-are/organization/bureaus/bureau-democracy-conflict-and-humanitarian-assistance/office-food

Wescott, Paul C. and Linwood A. Hoffman (1999). “Price Determination for Corn and Wheat: The Role of Market Factors and Government Programs.” Market and Trade Economics Division, Economic Research Service, U.S. Department of Agriculture. Technical Bulletin No. 1878.

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/

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.