Putting Quantitative Trade Models to the Test: Evidence from the Trump Tariffs*

Rodrigo Adão
Chicago Booth

Arnaud Costinot
MIT

Dave Donaldson
MIT

Abstract

An important goal of a large quantitative trade literature is to shed light on the economic consequences of changes in tariffs, from unilateral trade liberalization episodes to the formation of regional trade agreements and trade wars. To help assess and potentially strengthen the credibility of the predictions of quantitative trade models we propose a new testing procedure that is intuitive, easy to implement, and consistent with standard practices in the field. Critically, unlike standard goodness of fit tests, our test is applicable in the presence of unobserved shocks that may be larger in magnitude than tariff changes and rich enough to rationalize any dataset. The basic idea behind our test is to exploit tariff changes, or observable shifters of these, that are orthogonal to unobserved shocks, a standard exclusion restriction in the empirical trade literature. As an illustration of how our test can be combined with state-of-the-art quantitative and empirical work, we test the welfare consequences of the Trump tariffs predicted by Fajgelbaum et al. (2020).

*This version: June 2022. Author contacts: rodrigo.adao@chicagobooth.edu, costinot@mit.edu, and ddonald@mit.edu. We are grateful to Robin Li, Nimisha Gupta, John Sturm and Akash Thakkar for outstanding research assistance and to seminar audience at Berkeley, Chicago, CREI, Duke, and Zürich for helpful comments.

1 Introduction

Policy makers around the world have faced, and will continue to face, the choice to liberalize trade or not. They may choose to cut their tariffs unilaterally, like India and Brazil in the 1990s, form regional trade agreements, like the EU, NAFTA, and Mercosur, or join the World Trade Organization, like China in 2001. They may also choose to raise their tariffs, as the Trump administration did in 2018, or leave existing regional agreements, as the United Kingdom did in 2020 after 47 years in the EU. One of the raisons d'être of a large quantitative trade literature is to provide guidance about these various policy choices by producing counterfactual simulations of how a country's imports, exports, terms of trade, tariff revenues, and ultimately welfare may change if a given trade policy were to be implemented.

While there is no doubt that the numbers provided by these simulations fill a demand for intellectual inputs into salient policy discussions, there is much more debate about their empirical credibility, both among academic researchers and policy makers.[1] The goal of this paper is to help assess and potentially strengthen the empirical credibility of quantitative trade models by developing a new, intuitive, and easy to implement test of their counterfactual predictions that is fully consistent with standard estimation and simulation procedures in the field.

The starting point of our testing procedure is the same as in Kehoe et al. (1995), Kehoe (2005), and Kehoe et al. (2017). We propose to go back in time and compare, ex post, the model's predictions about what would happen if tariffs were to change to what actually happened when the tariffs did change. However, the key empirical challenge when doing so is that other shocks, beside tariff changes, will surely have occurred over the period of interest. Technology, preferences, and other policies inevitably vary over time. As a result, standard goodness of fit tests based on the correlation between model predictions and data, such as those considered in the aforementioned papers and a wider literature discussed below, will reject models not because these models fail to predict the structural response of the economy to tariff changes, which is what researchers and policy makers are interested in, but because much of the variation in the data derives from the structural response to other shocks, which counterfactual predictions are agnostic about.

In fact, if one takes the existence of other shocks to the extreme and saturates quantitative trade models with sufficient free parameters such that they exactly match the data at any point in time, it may seem that there is little hope in ever testing quantitative

[1] Dawkins et al. (2001) provide an early discussion of concerns about the predictions of applied general equilibrium models, ranging from unrealistic assumptions to a general lack of transparency.

trade models. One may then settle instead for "theory with numbers," a view articulated early on by Shoven and Whalley (1984) and Dawkins et al. (2001). According to this view, models are useful filters to look at the world and measure which forces, including changes in tariffs, may be more important drivers of observed changes in a country's imports, exports, terms of trade, and welfare, see e.g. Eaton et al. (2016) and Caliendo et al. (2019). This approach to quantitative work, with exactly identified shocks and little emphasis on testing a model's counterfactual predictions, has become the dominant one in the field of international trade and spatial economics more generally—see e.g. Costinot and Rodríguez-Clare (2014) and Redding and Rossi-Hansberg (2017) for surveys.

In this paper, we put forward an alternative to standard goodness of fit tests that explicitly recognizes that there may be other shocks beside tariff changes; that these other shocks may indeed be much larger in magnitude; and that they may be rich enough to rationalize any dataset, consistent with standard practices in the field. Our idea is to exploit, in a general equilibrium context, the same type of exclusion restrictions that empirical researchers have previously used to estimate partial equilibrium elasticities, namely that other shocks are independent of either tariff changes or observable shifters of these. Under the null that the model is correct and that the exclusion restriction is satisfied, an instrumental variables (IV) regression of observed changes on model-predicted changes, instrumented by the general equilibrium impact of tariff changes or shifters of these, should deliver a coefficient of one. This is the general hypothesis that we propose to test.

Section 2 formalizes the previous discussion and introduces our IV test. We provide general sufficient conditions under which our test is valid and how these conditions differ from those invoked in previous goodness of fit tests. Rather than ignoring non-tariff shocks entirely, or allowing them to enter the model in entirely unrestricted ways, we instead leverage the fact that in many episodes researchers have reasons to believe that at least part of the tariff variation is orthogonal to other shocks. This allows us to compare the true structural response to tariff changes to the corresponding response predicted by a researcher's model, up to the projection on the IV. Intuitively, while the true structural response to tariff changes is not directly observable, since the data is generated from the sum of the structural response to tariff changes and other shocks, the projection of the true structural response to tariff changes on the IV is, since the orthogonality assumption guarantees that the projection of the true structural response to other shocks is zero.

Section 3 illustrates the small sample properties of our IV test for hypothetical random tariff shocks. To prepare our empirical analysis, we focus on the quantitative model used by Fajgelbaum, Goldberg, Kennedy and Khandelwal (2020) (FGKK) to analyze the

welfare consequences of the Trump tariffs. The model features multiple regions, multiple sectors, and input-output linkages in the country of interest, but treats the rest of the world as a series of import demand and export supply curves. Thus, general equilibrium considerations refer to adjustments in prices and quantities in the country of interest, but abstract from endogenous movements in import demand and export supply curves, potentially due to adjustments in factor prices and income effects in the rest of the world. The key structural parameters entering the model are the elasticities of import demand and export supply, which we also take from FGKK. The other structural parameters—a set of shifters of preferences and technology across products and countries—are set so that the model exactly matches US trade and production data in 2017, the year before the Trump tariffs were imposed, consistent with FGKK's counterfactual analysis.

We conduct Monte-Carlo simulations, in which we randomly draw tariff and non-tariff shocks for a sequence of model economies, both when FGKK's model is correctly specified and when it is not. Since the hallmark of quantitative trade models is to shed light on the welfare consequences of tariff changes, we focus our test on the three outcome variables whose changes determine aggregate welfare, up to a first-order approximation: import prices, export prices, and import quantities. The first two determine whether a country's terms-of-trade improve or worsen, whereas the third determines the extent of the fiscal externality associated with import tariff revenues.[2]

For our first series of simulations, we assume that FGKK's model is correctly specified and use it to generate data as well as the counterfactual predictions to be tested. Given these hypothetical data and predictions, we then construct 95% confidence intervals for our IV estimator and show that we only reject the null that the model is correctly specified at a rate close to 5%, giving credence to the sufficient conditions required for our test to be satisfied in the context of the US economy. In contrast, the correlation between observed changes and predicted changes may be anywhere between 0 and 1 depending on the variance of the tariff shocks. For our second series of simulations, we assume instead that FGKK's model is misspecified and generate data from alternative models that may differ from FGKK's model because of different elasticities. In this case, we show that our test is able to reject the null that the model is well specified with a probability that increases with the extent of misspecification. This is again in contrast to the simple correlation test that may remain consistently close to one and may even increase as we depart from the

[2]Under the assumption that the social marginal utility of income is equalized across households, a standard envelope argument therefore implies that, up to a first-order approximation, the aggregate welfare impact of the tariff change is equal to the weighted sum (across goods) of the import and export (log-) price changes, weighted by the share of imports and exports in GDP, plus the weighted sum of changes in import quantities, weighted by the share of tariff revenue in GDP.

true model.

Section 4 turns to the consequences of the actual Trump tariffs. We again focus on the predictions of FGKK's model for import prices, export prices, and import quantities, but instead of generating data for hypothetical tariff and non-tariff shocks, we now use actual data on these three outcome variables, both before and after the Trump tariffs are implemented. We view FGKK's analysis as the start-of-the art in terms of combining theory and empirics to shed light on the welfare implications of changes in trade policy. First, the authors use a quasi-experimental research design to estimate import demand and export supply elasticities in a credible way. Second, they combine these parameters with a structural model to generate predictions about the overall effects of changes in trade policy. We propose to add a third step that uses the same quasi-experimental research design to compare—in a formal sense—the general equilibrium predictions of the model with what is observed in the data.[3] In line with FGKK's empirical analysis, the exclusion restriction upon which we build is that the changes in US tariffs, which were driven by a change in the preferences of the policymakers elected in 2016 relative to their predecessors (e.g. a heightened desire to hurt Chinese interests in particular sectors), are plausibly orthogonal to US economic conditions (e.g. productivity growth in the same sectors).

These tests show that the quantitative trade model we evaluate makes a range of predictions, about welfare-centric outcomes, some of which line up closely with the data and some of which appear quite at odds with it. But, ultimately, a version of our test that is both particularly demanding—emphasizing cross-sectoral variation that was not used in FGKK's model parameter estimation—and welfare-oriented—weighting observations according to their first-order contribution to aggregate welfare—yields an IV coefficient of 1.55 and a p-value of 0.31. This implies that FGKK's model under-predicts the average response of the US terms of trade and tariff revenues by a factor of $1/3$, but that our test does not reject, at standard levels, the null hypothesis that the model is correctly specified. By contrast, the correlation between model predictions and data in this context is never higher than 0.1 (that is, the R-squared obtained when regression one variable on the other would never exceed 0.01), which highlights the difference between the conditional (that is, IV-based) nature of the comparisons in our test and that in standard goodness of fit

[3]To estimate directly the causal impact of a nation's tariff policy on the price and quantity changes that enter aggregate welfare, without imposing restrictions on the nature of general equilibrium linkages, one would need access to a long panel with exogenous variation in the entire vector of tariffs, not just a one-time change in the vector tariffs, even in the best case scenario where tariff changes are randomly drawn across products. For this reason, the researcher's model plays an indispensable role in extrapolating from what can be estimated given available tariff variation to what is desired. As our simulations demonstrate, the researcher's maintained hypothesis that the model is correctly specified can still be tested, and rejected when appropriate, even with access to a single time-change in a nation's tariff policy.

exercises.

Related Literature

The impact of trade liberalization. There is a large empirical literature estimating the impact of changes in trade policy recently reviewed by Goldberg and Pavcnik (2016). The authors contrast structural work based on quantitative trade models whose "estimated effects [...] depend on the assumption of the underlying structural model and the consistency of the estimated behavioral parameters of demand, supply, and implied trade elasticities" and reduced-form work exploiting a quasi-experimental research design, which "depends less on specific functional form assumptions about the underlying demand, production, and market structure" and can be used "to estimate the direct causal effect of actual trade policy on the outcomes of interest" but "is not suited to evaluate welfare implications of actual trade policy changes or the overall effects of trade policy change, both of which require fully specified structural or quantitative models." Examples of reduced-form work estimating the direct causal effect of actual trade policy includes Attanasio et al. (2004) for Columbia; Topalova (2010) for India; McCaig (2011) for Vietnam; and Kovak (2013) for Brazil, among many others.[4] We view our IV test as a useful and yet simple "add-on" to the existing literature for researchers interested in combining quantitative structural work and reduced-form empirical work. After estimating using quasi-experimental variation and simulating using a quantitative model, we advocate testing the quantitative model's predictions for welfare-centric outcomes by leveraging the same quasi-experimental variation.

Testing trade models. In addition to the tests of the comparative static predictions of quantitative trade models in Kehoe et al. (1995), Kehoe (2005), and Kehoe et al. (2017), there is a long tradition of testing in the international trade literature that focuses on cross-sectional predictions and does not rely on specific functional form assumptions. Tests of the Heckscher-Ohlin-Vanek model (Bowen et al., 1987; Trefler, 1995; Davis and Weinstein, 2001) are classic examples. Costinot and Donaldson (2012) is another example focusing on the cross-sectional predictions of the Ricardian model. In these papers, under the null that the model is well specified, there should be no residuals, like in the aforementioned

[4] Kovak (2013) also develops a specific-factor model that offers micro-foundations for standard Bartik-style regressions of regional labor market exposure to tariff changes. Empirically, he finds that the response of wages to tariff changes observed in the data is smaller than that predicted by his model, which he interprets as consistent with the presence of internal migration and incomplete pass-through from import tariffs to domestic producer prices (from which his formal model abstracts).

tests of quantitative trade models. Under this assumption, "sign tests," "variance tests," or "slope tests" are all equally valid procedures. In contrast, we propose to weaken the assumption imposed on the structural residuals, which may be non-zero and have substantial variance. Under this weaker assumption, we show that an IV version of the slope test remains a valid testing procedure, whereas other tests are not.

Another paper related to our work is Lai and Trefler (2002). The authors focus on the implications of changes in tariffs in a multi-sector gravity model. They decompose bilateral trade flows into an income term, an expenditure term, and a price term that directly depends on tariffs. Over a twenty-year period, they show that the R-squared of a regression of observed changes in bilateral trade on changes in the price term (adjusted for the two other terms) is very low. While we share the same broad objective, our testing procedure differs in two important ways. First, our analysis focuses on the full impact of tariffs, not the impact that remains after conditioning out their impact on income and expenditure. Second, the starting point of our test is that changes in tariffs (or some shifters of these) are orthogonal to other economic shocks, not that changes in tariffs have higher variance than other shocks. In this respect, we also differ from Dingel and Tintelnot (2021) who propose to compare the fit of alternative quantitative spatial models using standard goodness of fit measures, as in Kehoe et al. (1995), Kehoe (2005), and Kehoe et al. (2017), but applied to changes in commuting rather than trade flows.

Among existing tests, our procedure is most closely related to the tests implemented in Adao et al. (2020) and Adao et al. (forthcoming). At a technical level, both papers start by log-linearizing their model and then compare the elasticity of labor market equilibrium variables with respect to exogenous foreign shocks observed in the data to the elasticity predicted by the log-linearized version of the model. This procedure is valid for economies in which both foreign shocks and other shocks are sufficiently small, so that the model deviates from its first-order approximation only by small amounts. In contrast, the test we propose here is globally valid, which removes any concern that a model may be rejected because the log-linear approximation of the model is poor, not because the fully non-linear model itself is misspecified; our procedure deploys first-order approximations in, and only in, the construction of the instrumental variables that enter our testing procedure. At a more substantive level, neither of these two papers studies the impact of changes in trade policy, a historically major area of applications of quantitative trade models, which is the main focus of our analysis.

Credible counterfactual analysis. Our interest in enhancing the credibility of counterfactual predictions relates to our earlier work, Adao et al. (2017). In that earlier work, we

have shown that in a neoclassical environment, the shape of factor demand—and only the shape of factor demand—determines counterfactual predictions. This result made it easier to ask whether the moments chosen for structural estimation are related to the economic relation of interest and to explore whether functional form assumptions rather than data drive particular results. Here, instead of focusing on the estimation of the demand for factor services, we go back to a standard estimation of quantitative trade models, but ask, once the structural parameters have been estimated using standard moment conditions, whether, for a given counterfactual question of interest such as the welfare impact of changes in tariffs, a given model appears to do too much violence to the data in a statistical sense that we make precise.

2 Putting Quantitative Trade Models to the Test

2.1 A Bird's-Eye View of Quantitative Trade Models

A quantitative trade model, like any economic model, imposes restrictions on the behavior of endogenous variables, typically prices and quantities, as a function of exogenous variables, typically preference and productivity shocks as well as various taxes. For ease of exposition, suppose further that this quantitative trade model is static, as is often the case in the literature.[5] Then in any given period t, we can describe it compactly as a mapping f such that

$$y_t = f(\tau_t, \epsilon_t), \tag{1}$$

where y_t denotes the vector of all endogenous variables, either quantities or prices; τ_t denotes the vector of taxes imposed at date t, which in all our later applications will be import tariffs; and ϵ_t denotes the vector of all other time-varying shocks. Different assumptions about preferences, technology, and market structure lead to different mappings or "reduced-form" f that summarize the general equilibrium effects of taxes and other shocks, τ_t and ϵ_t, according to the researcher's model.[6]

To state the obvious, the set of potential mappings f is *very* large. Even if one is only

[5] The general points that we make about testing do not depend on this assumption. Focusing on a static model, however, simplifies notation. Since we will focus on static models in all subsequent sections, we prefer to adopt simpler notation right away. For expositional purposes, and in line with the rest of our analysis, we also ignore issues related to multiplicity of equilibria in which the predictions of a quantitative trade models may be sets rather than points.

[6] The mapping f is the "reduced-form" of the model in a Cowles Commission sense: it solves for all the endogenous variables as a function of the exogenous variables τ_t and ϵ_t, the same way one can explicitly solve for price and quantity in partial equilibrium as a function of supply and demand shifters—the counterparts of τ_t and ϵ_t—rather than describe them as the implicit solution of supply and demand equations.

interested in the impact of tax changes, there is little hope of ever getting enough time series variation to trace $f(\cdot, \epsilon_t)$ out non-parametrically. The typical approach to obtain knowledge of f is therefore to start from a micro-founded model where consumers maximize utility, typically of the nested CES form, firms maximize their profits, with production functions also typically of the nested CES form, and markets clear, typically in a Walrasian fashion. Since knowledge of f acquired in this way relies at least in part on many untested assumptions, it is not clear why a given quantitative trade model would actually be a good approximation to the true data-generating process:

$$y_t = f^*(\tau_t, \epsilon_t^*). \tag{2}$$

The question that we are interested in is whether, despite the fact that f abstracts from many features of reality and invokes strong functional form assumptions, its counterfactual predictions about the impact of tariff changes, $\Delta x \equiv f(\tau_{t+1}, \epsilon_t) - f(\tau_t, \epsilon_t)$, are close, in some metric to be defined, to the true structural impact of tariff changes, $\Delta x^* \equiv f^*(\tau_{t+1}, \epsilon_t^*) - f(\tau_t, \epsilon_t^*)$.

2.2 The Potential Problem with Standard Goodness of Fit Tests

Perhaps the most intuitive way to test quantitative trade models is to compare their predictions, $\Delta x \equiv f(\tau_{t+1}, \epsilon_t) - f(\tau_t, \epsilon_t)$, to the changes observed in the data, $\Delta y \equiv y_{t+1} - y_t$. Under the null that a researcher's model is the true model, $f = f^*$, and that there are no other shocks, $\epsilon_t = \epsilon_{t+1}$, the two vectors Δx and Δy should be identical.[7] Therefore, one may: (i) compute the correlation between Δy and Δx; or (ii) compute the R-squared of a linear regression of Δy on Δx; or (iii) estimate the intercept and slope of the same linear regression. Under the joint hypothesis that $f = f^*$ and $\epsilon_t = \epsilon_{t+1}$, the quantitative trade model is correctly specified and that there are no other shocks, one should observe that: (i) the correlation is 1; (ii) the R-squared is 1; and (iii) the estimated intercept and slope are equal to 0 and 1, respectively. These observations are the basis of the various tests in Kehoe et al. (1995), Kehoe (2005), and Kehoe et al. (2017). They are similar in spirit to the goodness of fit tests considered by Davis and Weinstein (2001) in the Heckscher-Ohlin-Vanek literature. In both cases, since the model is supposed to hold exactly, any of the three metrics can be used to compare predictions and data.

[7] More generally, if ϵ_t includes shocks that are directly observable, $\epsilon_t \equiv (\epsilon_{\text{observed},t}, \epsilon_{\text{unobserved},t})$, then one may also compare Δy to $\Delta x \equiv f(\tau_{t+1}, \epsilon_{\text{observed},t+1}, \epsilon_{\text{unobserved},t}) - f(\tau_t, \epsilon_t)$, i.e., what the model predicts when only holding fixed the unobserved component of other shocks, i.e. $\epsilon_{\text{unobserved},t+1} = \epsilon_{\text{unobserved},t}$. This is the path taken by Kehoe et al. (1995) who, when considering the impact of Spain joining the European Community in 1986, also allow for observed changes in oil prices and agricultural productivity.

We refer to these exercises as standard goodness of fit tests. The potential problem with these tests is that they are not designed to distinguish whether $\Delta x \neq \Delta y$ arises because of $f \neq f^*$ or because $\epsilon_t \neq \epsilon_{t+1}$ is large. For a forecaster interested in predicting Δy, the distinction is irrelevant, since accurately predicting Δy is the object of interest. But we believe—and we think that many other researchers would agree—that the role of quantitative trade models is *not* to forecast future changes in economic variables. Their role is more modest; it is to offer *ceteris paribus* comparisons between economies of interest and counterfactual versions of those in which a particular policy change is or is not implemented. That is, Δx^* rather than Δy is what we would like acquire knowledge of. For instance, even if we were unable to predict real GDP growth in the US, because it depends on changes in productivity across sectors and various other shocks, we may still value precise information about how different real GDP growth in the US would have been absent the Trump tariffs. We now present an empirical test designed to draw that distinction.

2.3 This Paper's Empirical Test

Our empirical test builds on two basic observations. First, there are other non-tariff shocks whose magnitude may be large, even around well-known trade liberalization episodes, in line with standard practices in the quantitative trade and spatial literature (Costinot and Rodríguez-Clare, 2014; Redding and Rossi-Hansberg, 2017). Second, while we may not have strong priors about the magnitude of non-tariff shocks, we may be confident, depending on the particular setting, that these shocks are orthogonal to tariffs or some instrumental variable (IV) of tariffs, in line with standard practices in the empirical literature estimating the causal impact of tariffs (Goldberg and Pavcnik, 2016). The first observation implies that we cannot directly compare Δx to Δx^*, since it differs from Δy, whereas the second observation opens up the possibility of comparing the projections of Δx and Δx^* on the IV, since the latter coincides with the projection of Δy on the IV under the exclusion restriction.

An IV Test. Take a subset of endogenous variables $\{y_{n,t}\}_{n=1,...,N}$ included in y_t. In our applications, given our interest in testing the welfare predictions of quantitative trade models, these variables will be the sufficient statistics for computing changes in aggregate welfare, up to a first-order approximation, in the model to be tested: import prices, import quantities, export prices, or linear combinations of those, with $n = 1, ..., N$ the index of tradable goods.

Our empirical test requires the existence of an IV, $\{z_n\}_{n=1,\ldots,N}$, that is determined outside of equation (2), correlated with the impact of tariff predicted by the researcher's model, Δx_n, and orthogonal to the structural response to non-tariff shocks in the true model, which we denote $\Delta\eta_n^* \equiv f_n^*(\tau_{t+1}, \epsilon_{t+1}^*) - f_n^*(\tau_{t+1}, \epsilon_t^*)$.

A1 [Existence of IV]. *There exists* $\{z_n\}_{n=1,\ldots,N}$ *such that* (i) $plim_{N\to\infty} \frac{1}{N} \sum_n z_n \Delta x_n \neq 0$ *and* (ii) $plim_{N\to\infty} \frac{1}{N} \sum_n z_n \Delta\eta_n^* = 0$.

We discuss in detail below how to construct an IV that satisfies Assumption A1. Before doing so, we demonstrate how the existence of an IV can be used to design an alternative goodness-of-fit measure centered on the difference between the structural responses to tariff shock, Δx and Δx^*. Formally, let $\beta_{IV} \equiv \sum_n z_n \Delta y_n / \sum_n z_n \Delta x_n$ denote the IV estimator associated with the reduced-form and first-stage effects of the IV z_n on the observed changes Δy_n and the researcher's predictions Δx_n,

$$\Delta y_n = \beta_y z_n + \varepsilon_{y,n},$$
$$\Delta x_n = \beta_x z_n + \varepsilon_{x,n},$$

with $\beta_{IV} = \beta_y / \beta_x$. Part (i) of Assumption A1 states that the instrumental variable z_n is correlated with the predicted impact of tariff Δx_n. That is, the first stage of the IV procedure is non-zero, $\beta_x \neq 0$, which one can check directly in the data. Part (ii) of Assumption A1 states that the only reason for the reduced-form relationship between the instrumental variable z_n and the observed change in the outcome of interest $\Delta y_n = \Delta x_n^* + \Delta\eta_n^*$ is the true structural response, Δx_n^*. Intuitively, we can therefore test whether $\Delta x_n^* = \Delta x_n$ by comparing whether Δy_n and Δx_n covary in the same way with the IV z_n. Proposition 1 formalizes this intuition.

Proposition 1. *Suppose that A1 holds, then*

$$plim_{N\to\infty} (\beta_{IV} - 1) = plim_{N\to\infty} \sum_n \left(\frac{z_n \Delta x_n}{\sum_m z_m \Delta x_m} \right) \left(\frac{\Delta x_n^* - \Delta x_n}{\Delta x_n} \right).$$

Proof. Start from $\beta_{IV} \equiv \frac{\sum_n z_n \Delta y_n}{\sum_n z_n \Delta x_n}$. Substitute Δy_n using equation (2) to get $\beta_{IV} - 1 = \frac{\sum_n z_n [\Delta x_n^* + \Delta\eta_n^* - \Delta x_n]}{\sum_n z_n \Delta x_n}$. Then take the limit of the previous expression and invoke Assumption A1 to obtain $\text{plim}_{N\to\infty} \beta_{IV} - 1 = \text{plim}_{N\to\infty} \sum_n \left[\frac{z_n \Delta x_n}{\sum_m z_m \Delta x_m} \right] \left[\frac{\Delta x_n^* - \Delta x_n}{\Delta x_n} \right]$. □

Proposition 1 opens up the possibility of testing quantitative trade models by testing whether β_{IV} is equal to one against the alternative that it is not. If the researcher's model is the true model, $f = f^*$, and Assumption A1 holds, Proposition 1 implies that

$\text{plim}_{N\to\infty}\beta_{IV} = 1$, since $\Delta x_n^* = \Delta x_n$ for all n. Conversely, if the researcher's model is distinct from the true model, $f \neq f^*$, and Assumption A1 holds, then $\text{plim}_{N\to\infty}\beta_{IV} \neq 1$ provided that the true model generates different structural responses for the outcome variables of interest, $\Delta x_n \neq \Delta x_n^*$, and that these different responses are correlated with the IV, so that $\sum_n z_n[\Delta x_n^* - \Delta x_n] \neq 0$. This final observation is important; it implies that the choice of the IV matters for the performance of our test. Indeed, Proposition 1 formally establishes that the difference between β_{IV} and 1 is a consistent estimator of the average difference between the true structural response, Δx_n^*, and the researcher's predicted response, Δx_n, but with weights determined by the projection of Δx_n on the IV z_n.

Critically, the previous "IV test" does not rely on any assumption that tariff changes are the only changes, $\Delta\eta_n^* = 0$, or that they are large relative to other shocks in the sense that $\text{var}(\Delta\eta_n^*) \ll \text{var}(\Delta x_n)$. This is in contrast to goodness of fit tests based on correlation or R-squared discussed earlier. Even in the best case scenario in which the model is right and tariff shocks are uncorrelated with other shocks—i.e. $\beta = 1$ and $\text{plim}_{N\to\infty}\frac{1}{N}\sum_n \Delta x_n \Delta\eta_n^* = 0$—the R-squared of the OLS regression of Δy_n on Δx_n is $R^2 = \left(1 + \frac{\text{var}(\Delta\eta_n^*)}{\text{var}(\Delta x_n)}\right)^{-1}$, which differs from one if $\text{var}(\Delta\eta_n^*) \neq 0$ and actually converges to zero as $\text{var}(\Delta\eta_n^*)/\text{var}(\Delta x_n)$ goes to infinity. Since $R^2 = (\text{corr}(\Delta y_n, \Delta x_n))^2$, the same observation implies that even the predictions of a true model may have a vanishingly small correlation with what is observed in the data.

Construction of an IV and Asymptotic Distribution of β_{IV}. To construct an IV, we have two broad goals in mind. The first one is standard: z_n should be designed such that the exclusion restriction in Assumption A1 is credible. The second one is more unique to our testing procedure: conditional on A1 holding, z_n should also be designed such that the weights $z_n\Delta x_n / \sum_m z_m \Delta x_m$ appearing in Proposition 1 are aligned with the counterfactual question of interest. For instance, when considering changes in import and export prices, one would like an instrument z_n that correlates well with the initial value of imports and exports, consistent with a first-order approximation to welfare changes, and likewise, when considering changes in import quantities, one may want to weigh the instrument to correlate with initial tariff revenue. For any of the IV used in this paper, we propose to start from a first-order approximation to the predicted impact of the tariff change, $\sum_m (\partial f_n/\partial\tau_m)\Delta\tau_m$, and substitute in for $\Delta\tau = \tau_{t+1} - \tau_t$ by using an exogenous shifter of tariff that is observable to the researcher, $\Delta\tau_{IV}$.[8] That is, we use

[8]Depending on the context, the shifter may be the change in tariff itself, $\Delta\tau_{IV} = \Delta\tau$, as in Fajgelbaum et al. (2020), or (the negative of) the initial level of the tariff $\Delta\tau_{IV} = -\tau_t$, as in Topalova (2010).

$\sum_m (\partial f_n/\partial \tau_m)\Delta\tau_{IV,m}$ as the predicted impact of the tariff change for good n. This procedure aims to combine existing exclusion restrictions from the empirical literature, as reflected in the choice of $\Delta\tau_{IV}$, with the general equilibrium structure of our model, as summarized by the gradient, $\{\partial f_n/\partial \tau_m\}$. In our applications we will use the demeaned version $z_n \equiv \sum_m (\partial f_n/\partial \tau_m)(\Delta\tau_{IV,m} - \Delta\bar{\tau}_{IV})$ as our baseline IV, where $\Delta\bar{\tau}_{IV}$ is a weighted average of the initial tariffs, $\Delta\bar{\tau}_{IV} \equiv \frac{1}{M}\sum_m \omega_m \Delta\tau_{IV,m}$, and the weights ω_m are chosen so that both the expectation and sample-mean of z_n is zero. We will then consider variants of this IV that satisfy A1 under the same primitive assumptions (to be described below), but which put different weights on different sources of misspecification, $(\Delta x_n^* - \Delta x_n)/\Delta x_n$.

The mathematical structure of our IV is similar to that of a "shift-share IV," with the gradient $\{\partial f_n/\partial \tau_m\}$ playing the role of the "shares" and the vector of initial tariffs τ_t playing the role of the "shifts." This is not innocuous. Because of general equilibrium linkages, neither the changes in the variables of interest Δy_n nor the researcher's prediction Δx_n are i.i.d. across goods n. Rather they are n-specific functions of the same underlying change in the vector of tariffs and other shocks. This raises non-trivial questions about which law of large numbers to invoke for the consistency of β_{IV} and which central limit theorem to turn to for computing its standard error, a critical step for testing.

The shift-share structure of our IV allows us to focus on a special case with non i.i.d. variables for which such results already exist. To establish that Assumption A1 holds and establish the consistency of β_{IV}, we can use the results of Borusyak et al. (2022) and provide primitive restrictions on the data generating process for$(\Delta\tau_{IV}, \tau_{t+1}, \epsilon_t, \epsilon_{t+1})$ such that the exclusion restriction $\text{plim}_{N\to\infty}\frac{1}{N}\sum_n z_n \Delta\eta_n^* = 0$ is satisfied, as shown in Appendix A.1. Likewise, to construct the confidence interval of β_{IV}, we can use the results of Adao et al. (2019). Proposition 2, whose proof can be found in Appendix A.2, describes the asymptotic distribution of β_{IV} under the null that the researcher's model is the true model.

Proposition 2. *Suppose that* (i) $\Delta\tau_{IV,m}$ *are i.i.d across* m *and independent of* $(\tau_{t+1}, \epsilon_t, \epsilon_{t+1})$; (ii) $\frac{1}{N^2}\sum_m (\alpha_{m,t})^2 \to 0$ *and* $\frac{\max_m(\alpha_{m,t})}{\sum_r \alpha_{r,t}^2} \to 0$, *with* $\alpha_{m,t} \equiv \sum_n [(\partial f_n/\partial \tau_m) - \omega_m \sum_r (\partial f_n/\partial \tau_r)]$; *and* (iii) $Var[\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = \sigma^2 < \infty$ *and* $E[(\Delta\tau_{IV,m})^4|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \leq T$ *for all* m *and* $\Delta\eta_n^* \leq M$ *for all* n. *Then if* $f = f^*$,

$$\frac{\sum_n z_n \Delta x_n}{\left(\sum_m (\alpha_{m,t})^2\right)^{1/2}}(\beta_{IV} - 1) \to_d \mathcal{N}(0, V_N)$$

with the variance $V_N \equiv \sum_m \{\sigma \sum_n [(\partial f_n/\partial \tau_m) - \omega_m \sum_r (\partial f_n/\partial \tau_r)]\Delta\eta_n^*\}^2 \Big/ \sum_m (\alpha_{m,t})^2$.

Condition (i) implies that while the IVs z_n are not i.i.d across n, they can be expressed as (n-specific) linear combinations of i.i.d variables $\Delta\tau_{IV,m}$ whose realizations are inde-

pendent of the structural residual $\Delta\eta_n^*$.[9] Provided that these linear combinations do not all tend to load on the same tariff shifters—condition (ii)—and that standard regularity conditions hold—condition (iii)—the previous assumptions provide the random variation required to establish law-of-large numbers, so that $\text{plim}_{N\to\infty}\frac{1}{N}\sum_n z_n\Delta\eta_n^* = 0$ and $\text{plim}_{N\to\infty}\beta_{IV} = 1$ if $f = f^*$ by Proposition 1, as well as central limit theorems, so that $\beta_{IV} - 1$, appropriately scaled, is normally distributed, as described in Proposition 2.[10] These are the formal results that we will use to test the predictions of quantitative trade models in the rest of this paper.

2.4 Estimate, then Test!

Leamer and Levinsohn (1995) famously exhorted trade economists to "Estimate, don't test!". Our paper is unapologetically about testing. However, we do not view estimation and testing as mutually exclusive alternatives. In fact, before testing the general equilibrium implications of a given quantitative trade model, we will rely on estimates of its key structural parameters. To help clarify some of the issues that arise when both estimation and testing are conducted, suppose that the reduced-form of the quantitative model of interest can be decomposed into

$$f(\tau_t, \epsilon_t) \equiv g(\tau_t, \epsilon_t|\hat{\theta}),$$

where $\hat{\theta}$ is the estimator of a structural parameter obtained from some generic moment condition, $\frac{1}{N}\sum_n h_n(y_t, y_{t+1}, \tau_t, \tau_{t+1}, \Delta\tau_{IV}; \theta) = 0$. The mapping g, in turn, reflects all other assumptions in the model that do not derive from estimation, for instance that some groups of factors and goods are perfect substitutes, that input-output linkages are Cobb-Douglas, or that markets are perfectly or monopolistically competitive.

Let us start with the obvious. Provided that the moment condition used for estimation, $\frac{1}{N}\sum_n h_n(y_t, y_{t+1}, \tau_t, \tau_{t+1}, \Delta\tau_{IV}; \theta) = 0$, is distinct from the moment used for testing,

[9] The fact that tariffs in the post-period $t+1$ also appear in condition (i) reflects the fact that the model that we are testing is potentially non-linear. To see this formally, note that if $f = f^*$ then, up to a first-order approximation, one can write $\Delta y_n = \Delta^{\text{f.o.a}} x_n + \Delta^{\text{f.o.a}}\eta_n^*$, with the two first order terms given by $\Delta^{\text{f.o.a}} x_n = \sum_m (\partial f_n/\partial\tau_m)|_{(\tau_t,\epsilon_t)}\Delta\tau_m$ and $\Delta^{\text{f.o.a}}\eta_n^* = \sum_m (\partial f_n/\partial\epsilon_m)|_{(\tau_t,\epsilon_t)}\Delta\epsilon_m$, so that the structural residual is independent of τ_{t+1}.

[10] Following Adao et al. (2019), we obtain an estimator of the variance of $\beta_{IV} - 1$ in Proposition 2 by substituting the structural residual by its finite sample analog, $\Delta\hat{\eta}_n^* = \Delta y_n - \beta_{IV}\Delta x_n$, leading to

$$\hat{V}_{AKM} = \frac{\sum_m \{\sigma \sum_n [(\partial f_n/\partial\tau_m) - \omega_m \sum_r (\partial f_n/\partial\tau_r)][\Delta y_n - \beta_{IV}\Delta x_n]\}^2}{\left(\sum_n z_n \Delta x_n\right)^2}.$$

$\frac{1}{N}\sum_n z_n \Delta\eta_n^* = 0$, one can both estimate and test.[11] In particular, one may focus on the same variables for the purposes of estimation and testing. The more subtle observation that motivates our analysis is that although the two previous moments conditions are distinct, they may derive from the same primitive restrictions on the data generating process, for instance that shifters of tariffs are independent of other shocks affecting the economy of interest, as in Proposition 2. Thus if one is already confident about these primitive restrictions for the purposes of estimation, as in many existing empirical papers, one should be equally confident about testing quantitative trade models using these restrictions.

A second issue is the relationship between our test and the standard practice in quantitative work of using a subset of "targeted moments" for estimation and another subset of "untargeted moments" as a way to build confidence in the model of interest.[12] At a general level, one can think of our IV test as the second step of such a general procedure. A common issue with such a procedure, though, is the choice of targeted versus untargeted moments. If all moments are obtained from a given cross-section, why would one moment belong to the second group rather than the first? A distinct feature of our test is that it is directly motivated by a specific question, namely predicting the welfare impact of tariff changes. This motivation suggests selecting untargeted moments that give center stage to the observed responses of the endogenous variables that shape welfare to changes in tariffs, since these are precisely the dimensions along which we would like to be more confident about, and using an IV to purge these observed responses from the contribution of other shock.[13]

A final important question relates to the dimension of the model that is being put to the test by a given procedure. Models are simplifications and there are myriad of ways

[11] In practice, very few trade papers use the reduced-form of their model for the purposes of estimation. The only exception that we are aware of is Adao et al. (2020), which uses the reduced-form impact of the China shock on employment and wages to estimate labor supply and demand across US commuting zones.

[12] While we refer to this practice as standard, we note that it is much more standard in IO and macro than in international trade. As mentioned above, recent quantitative trade models often feature shocks that are flexible enough to match the data perfectly and hence, trivially, to match any moment that one could compute from such data. For instance, in their general-equilibrium analysis of the China shock, Caliendo et al. (2019) note that, "since our baseline economy matches the factual economy, if we run the second stage of the ADH regression in our baseline economy, by construction we will replicate the ADH regression results." Under this approach, empirical evidence on the observed labor market consequences of the China shock, such as those provided by Autor et al. (2013), can be used as motivation, but never as an input into a testing procedure.

[13] An alternative to our two-step approach would be to use the two moments, $\frac{1}{N}\sum_n h_n(y_t, y_{t+1}, \tau_t, \tau_{t+1}, \Delta\tau_{IV}; \theta) = 0$ and $\frac{1}{N}\sum_n z_n \Delta\eta_n^* = 0$, for estimation in a stacked GMM procedure and then perform a J-test. For those who value the transparency of existing estimation procedures (for instance, by directly estimating import demand curves by tracing out how quantities respond to supply-side variation in prices), we see value in proposing a test that does not require changing these procedures (in that same example, by asking researchers to adjust import demand elasticities to match the general-equilibrium response of import prices and quantities).

in which quantitative trade models may be misspecified. We have already discussed how the choice of different IVs (that all satisfy Assumption A1) allows us to put different weights on different sources of mispecifications and, in turn, better align those with the counterfactual question of interest. By choosing different IVs, our test also allows us to focus attention on parts of the model that have not been used for estimation, for instance by taking sector-level averages of our baseline IV that only vary between broad sector categories when all estimation had been conducted within these categories.

We conclude by noting that the fact that our instrumental variable z_n is demeaned is not innocuous for the previous discussion. While demeaning is important to guarantee that Assumption A1 holds (since we do not want to impose that $\Delta\eta_n$ is mean zero), it implies that our test cannot discriminate between models that generate the exact same predictions, up to a constant. To see this, note that if $\Delta x_n = \Delta x_n^* + \text{constant}$, then $\text{plim}_{N\to\infty} \frac{1}{N} \sum_n z_n [\Delta x_n^* - \Delta x_n] = \text{constant} \times \text{plim}_{N\to\infty} \frac{1}{N} \sum_n z_n$.[14]

3 IV Test Results—using Simulated Tariff Changes

We now explore the properties of our test through a series of Monte-Carlo simulations in which we control the true data generating process. The researcher's model that we propose to test is the quantitative trade model developed by Fajgelbaum et al. (2020) (FGKK) to analyze the impact of the Trump tariffs on the US economy. This serves the dual role of understanding the properties of our test in a realistic setting and a study of the finite-sample properties of our actual test in Section 4. The only difference between the tests implemented here and those in Section 4 relates to the structure of the shocks. Here, we subject the model US economy to a sequence of hypothetical changes in tariffs and other shocks that are randomly drawn, both when the model is correctly specified and when it is not. In Section 4, we will repeat the same exercise, but feed in the actual tariff shocks experienced during the 2017-19 period.

3.1 The Researcher's Model: Fajgelbaum et al. (2020)

Consider a world economy comprising Home, indexed by $i = H$, and its trading partners, indexed by $i \neq H$. Home is composed of many regions, indexed by r, whose domestic households can be employed in one of many sectors, indexed by s. Time is discrete and

[14] It is important to note, however, that the issue is not the standard "intercept" issue, i.e. that a difference-in-difference estimator only identifies relative changes in a variable of interest. Our preferred regression will be pooling export prices and import prices. Missing a common intercept for both variables has zero implication for aggregate welfare, which is the variable we are ultimately interested in.

indexed by t. We let $L_{rs,t}$ denote the inelastic supply of labor to sector s in region r at date t and $w_{rs,t}$ the wage rate of households employed in that sector and region.

Domestic Preferences. All domestic households have the same nested CES preferences over products, indexed by g, from different sectors and countries,

$$U_t = (C_{NT,t})^{\beta_{NT,t}} (C_{T,t})^{1-\beta_{NT,t}}, \tag{3}$$

$$C_{T,t} = \prod_{s \in \mathcal{S}} (C_{Ts,t})^{B_{s,t}}, \tag{4}$$

$$C_{Ts,t} = \left[(A_{Ds,t})^{\frac{1}{\kappa}} (D_{s,t})^{\frac{\kappa-1}{\kappa}} + (A_{Ms,t})^{\frac{1}{\kappa}} (D^*_{s,t})^{\frac{\kappa-1}{\kappa}} \right]^{\frac{\kappa}{\kappa-1}}, \tag{5}$$

$$D_{s,t} = \left[\sum_{g \in \mathcal{G}_s} (a_{Dg,t})^{\frac{1}{\eta}} (d_{Hg,t})^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}, \tag{6}$$

$$D^*_{s,t} = \left[\sum_{g \in \mathcal{G}_s} (a_{D^*g,t})^{\frac{1}{\eta}} (d^*_{g,t})^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}, \tag{7}$$

$$d^*_{g,t} = \left[\sum_{i \neq H} (a_{ig,t})^{\frac{1}{\sigma}} (d_{ig,t})^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}, \tag{8}$$

where $\mathcal{S}$ is the set of all tradable sectors; $\mathcal{G}_s$ is the set of all products from sector s; $\beta_{NT,t}$, $\{B_{s,t}, A_{Ds,t}, A_{D^*s,t}\}_{s \in \mathcal{S}}$, $a_{Dg,t}$, $a_{D^*g,t}$, and $\{a_{ig,t}\}_{i \neq H}$ are exogenous preference shifters; $\kappa \geq 0$ is the elasticity of substitution between domestic consumption, $D_{s,t}$, and imports, $D^*_{s,t}$, within a given sector s; $\eta \geq 0$ is the lower-level elasticity of substitution between products, within each one of the previous nests; $\sigma \geq 0$ is the elasticity of substitution between foreign sources, within a given product; and. Following standard practices, we refer to the combination of a product g and an origin i in equation (8) as a "variety".

Domestic Technology. In each region r and sector s, non-tradable goods are produced one-to-on from labor,

$$Q_{NTr,t} = Z_{NTr,t} L_{NTr,t}, \tag{9}$$

whereas tradable goods are produced according to

$$Q_{sr,t} = Z_{sr,t} \left(I_{sr,t} \right)^{\alpha_{Is,t}} \left(L_{rs,t} \right)^{\alpha_{Ls,t}}, \tag{10}$$

$$I_{sr,t} = \prod_{k \in \mathcal{S}} (I_{ksr,t})^{\alpha_{ks,t}}, \tag{11}$$

with $\alpha_{Is,t} + \alpha_{Ls,t} \leq 1$ and $\sum_{k \in \mathcal{S}} \alpha_{ks,t} = 1$. Tradable intermediates, $I_{ksr,t}$, from sector k demanded by sector s and region r are produced in the exact same nested CES way as final consumption from that sector, $C_{Tk,t}$, as described by equations (5)-(8). For future reference, we let $q_{igsr,t}$ denote the quantity of variety ig used by firms from sector s in region r. Finally, given sector-level output, $\sum_r Q_{sr,t}$, the vector of product-level output, $\{q_{gs,t}\}_{g \in \mathcal{G}_s}$, satisfies

$$\sum_{g \in \mathcal{G}_s} q_{gs,t}/z_{gs,t} = \sum_r Q_{sr,t}, \tag{12}$$

Like the preference shifters above, $Z_{NTr,t}$, $Z_{sr,t}$, and $z_{gs,t}$ are exogenous and potentially time-varying productivity shifters.

Prices, Import Tariffs, and Transfers. There are no domestic transportation costs so prices of tradables are equalized across regions at Home. For any variety imported by Home, we let $p_{ig,t}$ denote the price of product g from country $i \neq H$ faced by domestic households and firms. Likewise, for any variety exported by Home, we let $p^X_{ig,t}$ denote the price received by domestic firms, with the convention that $p_{Hg,t} = p^X_{Hg,t}$ for domestic goods sold domestically. Home's ad-valorem import tariff $\tau_{ig,t}$ drives a wedge between the domestic import price $p_{ig,t}$ and the price received by firms from country i, $p^*_{ig,t}$, whereas foreign tariffs drive $\tau^*_{ig,t}$ a wedge between the domestic export price $p^X_{ig,t}$ and the price paid by households and firms from country i, $p^{X,*}_{ig,t}$

$$p_{ig,t} = (1 + \tau_{ig,t}) p^*_{ig,t}, \tag{13}$$

$$p^{X,*}_{ig,t} = (1 + \tau^*_{ig,t}) p^X_{Hg,t}. \tag{14}$$

There are no other taxes. The government at Home rebates total tariff revenues through a lump-sum transfer, T_t. Its budget constraint can be expressed as

$$T_t = \sum_{s \in \mathcal{S}, g \in \mathcal{G}_s, i \in \mathcal{I}} \frac{\tau_{ig,t}}{1 + \tau_{ig,t}} p_{ig,t} m_{ig,t} + D_t. \tag{15}$$

where $m_{ig,t} = d_{ig,t} + \sum_{r \in \mathcal{R}, s \in \mathcal{S}} q_{igsrt,t}$ denotes total imports of variety ig by Home.

Foreign Import Demand and Export Supply. The rest of the world is modeled as a series of import demand and export supply curves that determine the quantities $x_{ig,t}$ and

$m_{ig,t}$ of any product g exported and imported, respectively, by Home to any country i,

$$x_{ig,t} = a^*_{ig,t}(p^{X,*}_{ig,t})^{-\sigma^*}, \tag{16}$$

$$m_{ig,t} = (z^*_{ig,t})^{\frac{1}{\omega^*}}(p^*_{ig,t})^{\frac{1}{\omega^*}}, \tag{17}$$

where $a^*_{ig,t}$ and $z^*_{ig,t}$ are exogenous import demand and export supply shifters; σ^* is the elasticity of foreign import demand; and ω^* is the inverse of the elasticity of foreign export supply.

Competitive Equilibrium. For a given policy vector, $\tau_t \equiv \{\tau_{ig,t}, \tau^*_{ig,t}, T_t\}$, a competitive equilibrium corresponds to a vector of prices $p_t \equiv \{p_{ig,t}, p^X_{ig,t}, p^*_{ig,t}, p^{X,*}_{ig,t}, w_{rs,t}\}$ and quantities $q_t \equiv \{d_{ig,t}, q_{igrs,t}, m_{ig,t}, x_{ig,t}, \ell_{rs,t}\}$ such that domestic households maximize their utility, as described in (3)-(8), subject to their budget constraint; domestic firms maximize their profits subject to technological constraints, as described in (9)-(12); import and export prices satisfy the non-arbitrage conditions (13) and (14); the domestic government's budget is balanced, as described in (15); foreigners are on their export supply and import demand curves, as described in (16) and (17); and goods and labor markets at Home clear.

3.2 Calibration of the Researcher's Model

Given a vector of time-invariant elasticities and a vector of time-varying shocks,

$$\Theta \equiv \{\kappa, \eta, \sigma, \sigma^*, \omega^*\}$$

$$\epsilon_t \equiv \{\beta_{NT,t}, B_{s,t}, A_{Ds,t}, A_{Ms,t}, a_{Dg,t}, a_{Mg,t}, a_{ig,t}, Z_{NTr,t}, Z_{sr,t}, \alpha_{Is,t}, \alpha_{Ls,t}, \alpha_{ks,t}, z_{grs,t}, a^*_{ig,t}, z^*_{ig,t}, D_t, L_{sr,t}\},$$

the previous equilibrium conditions determine the mapping $f(\cdot, \epsilon_t)$ from the vector of policy τ_t into endogenous prices and quantities, $y_t \equiv (p_t, q_t)$. Before proceeding to our simulations, we describe how we calibrate Θ and ϵ_t.

Demand and Supply Elasticities (Θ). We use the five elasticity estimates obtained by FGKK when using changes in US tariffs, and foreign retaliatory tariffs, in 2018. The first three such elasticities describe Home households' and firms' elasticities of substitution across domestic and foreign bundles goods within each tradable sector ($\kappa = 1.19$), across different products within the domestic or foreign bundles ($\eta = 1.53$), and across different foreign origins within the foreign bundle ($\sigma = 2.53$). The remaining two elasticities capture foreign firms' and households' elasticities of demand for imports sourced from

Home ($\sigma^* = 1.04$) and foreign firms' (inverse) elasticities of supply for exports to Home ($\omega^* = -0.002$).

Time-Varying Shocks (ϵ_t). In line with our empirical application below, we calibrate ϵ_t such that, given the estimated supply and demand elasticities $\Theta \equiv (\kappa, \eta, \sigma, \sigma^*, \omega^*)$ and the initial period tariffs τ_t, the model $f(\tau_t, \epsilon_t)$ exactly matches trade and production data from the US in 2017. All data are taken from the replication and data materials in FGKK. It comprises: *(i)* variety-level quantities and values for both exports and imports, for 71 foreign countries i (spanning 99% of US trade) and 10,228 tradable products g (based on 10-digit HS codes); *(ii)* sector-level revenues and expenditures on inputs of labor and intermediates from each sector, for 88 tradable sectors s (based on NAICS classifications); and *(iii)* county-sector employment.[15]

3.3 Testing, When the Researcher's Model is Correct

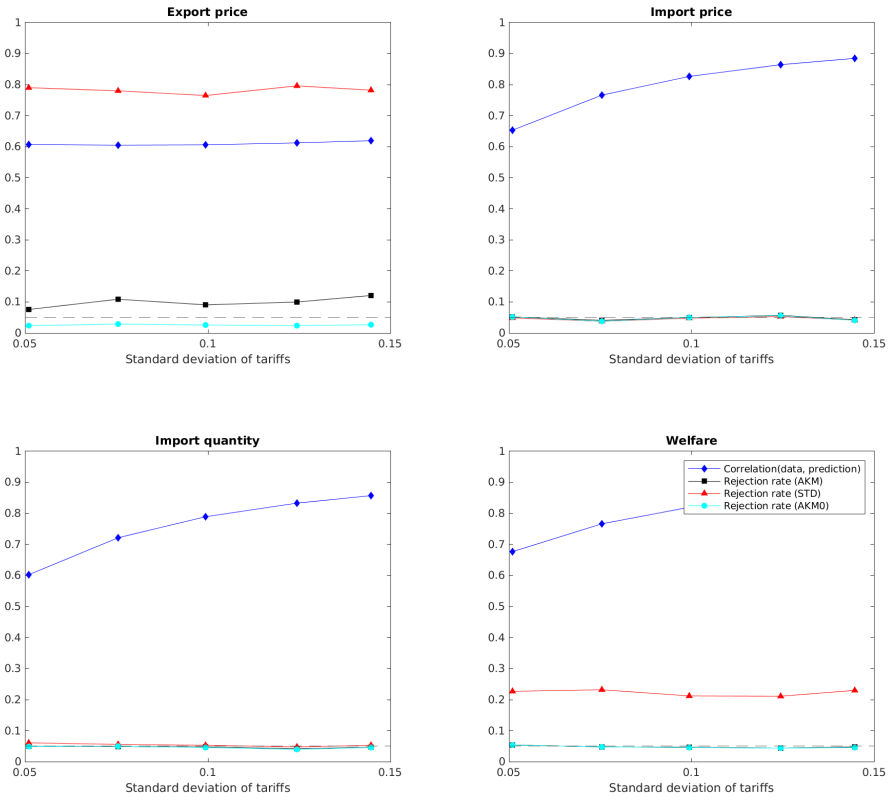
We begin by evaluating the properties of our test under the assumption that the researcher's model is the true data-generating process. We consider a sequence of 1,000 hypothetical economies, indexed by the superscript $b = 1, ..., 1000$, in which endogenous prices and quantities are generated by the equilibrium of the model in Section 3.1. For each hypothetical economy, we proceed in six steps.

First, we randomly draw initial domestic and foreign import tariffs $\{\tau^b_{g,t}, \tau^{*b}_{g,t}\}$ independently from a log-normal distribution (with mean zero and standard deviation that varies across simulations, as described below). Second, we use the procedure described in Section (3.2) to calibrate ϵ^b_t, given FGKK's estimates of the elasticities $(\kappa, \eta, \sigma, \sigma^*, \omega^*)$. Third, we randomly draw future tariffs $\{\tau^b_{g,t+1}, \tau^{*b}_{g,t+1}\}$ from the same log-normal distribution as in Step 1. Fourth, we randomly draw future values of non-tariff shocks, $\{\Delta a^*_{ig}, \Delta z^*_{ig}, \Delta a_{ig}, \Delta a_{D^*g}, \Delta Z_{rs}\}^b$, from normal distributions (with mean zero and standard deviation of 0.06). Fifth, we solve for changes in equilibrium outcomes Δy^b_n that result from the changes in tariffs and shocks of steps 3 and 4. In particular, we look at three outcomes: the change in the (log) prices of imports and exports, $\{\ln p_{ig,t}\}_{g,i\neq H}$ and $\{\ln p^X_{ig,t}\}_{g,i\neq H}$, and the change in the (log) quantity of imports, $\{\ln m_{ig,t}\}_{g,i\neq H}$; these three outcomes, when appropriately weighted, allow a first-order approximation to the effect of any tariff change on aggregate welfare. Sixth, we implement the test described in Section 2 with a 5% significance level where the predicted impact of the simulated tariff shock

[15] Specifically, for data that are observed annually (output, input use, and employment) we use the 2017 observation. But for the trade flow data, which are observed at monthly frequency, we use an average of the observations in the first four months of 2017.

Δx_n^b and the IV built from the simulated initial tariffs $z_{n,t}^b$ are computed using FGKK's model and its first-order approximation, respectively. We do so for each of the three outcomes separately as well as for the three outcomes simultaneously in a pooled regression. To compare our IV test to more traditional goodness of fit tests, we also compute the correlation between Δy_n^b and Δx_n^b.

Figure 1: Testing when the Researcher's Model is Correct



Notes: This figure reports the rejection rate of the test introduced in Section 2 for Δx_n^b given a 5% significance level (black squares), the equivalent rejection rate based on conventional standard errors (red triangles), the equivalent rejection rate based on AKM's conservative approach to testing under the null (turquoise circles), and the average correlation between Δy_n^b and Δx_n^b (blue diamonds) across $b = 1, ..., 1000$ simulated economies. We generate Δy_n^b and Δx_n^b by calibrating the model to match data for the US in 2017. The horizonal axis indicates the value of the standard deviation of the draws of import tariffs used in the simulations. Each panel reports results for the (log of the) outcome indicated in its title, with the bottom-right panel pooling all three outcomes.

Figure 1 reports the rejection rate of our test (black squares) and the average correlation (blue diamonds) across the 1,000 economies that we simulate for five different levels of the dispersion in the draws of import tariffs across sectors (used in steps 1 and 3). The blue diamonds show that, despite the model being correct in all simulations, the correlation between Δy_n^b and Δx_n^b varies substantially both across outcomes and, within outcomes, with the standard deviation of import tariffs. Not surprisingly, when the extent

of variation in import tariffs in a simulated data is low, the correlation tends to be correspondingly low; this simply indicates that any signal contained in the model's predictions is being swamped by other shocks in the model that affect the outcome of interest. It is noteworthy that the correlation is always lower for export-based outcomes than it is for import-based ones. This is a consequence of the fact that, given the variation in other shocks, domestic import tariffs have a direct impact on imports, but they only affect exports indirectly through their impact on wages and other domestic input prices. These latter forces are inherently more dispersed across sectors, which weakens the model's cross-sector explanatory power.

In contrast, the black squares show that the rejection rate of our test is largely invariant to how important import tariff shocks are relative to other shocks, with a median estimated coefficient extremely close to one in all cases. Our test rejects the true model at a rate of roughly 5% for import prices and quantities (and the "welfare" combination of all three outcomes reported in the lower-right figure) and at a rate of about 10% for export prices. Thus, the size of the test is correct for imports and the pooled regression, but the test slightly over-rejects the true model for exports. As discussed by Adao et al. (2019), these levels of over-rejection may arise when there are few sectors in practice. We therefore also report (in the turquoise circles) the results of using a conservative inferential procedure proposed by these authors.[16] Notably, however, a standard approach to inference (red triangles) would have strikingly incorrect rejection rates for some outcomes—for example, as high as 0.8 in the case of export prices—regardless of the variance of the simulated tariffs.

3.4 Testing, When the Researcher's Model is Incorrect

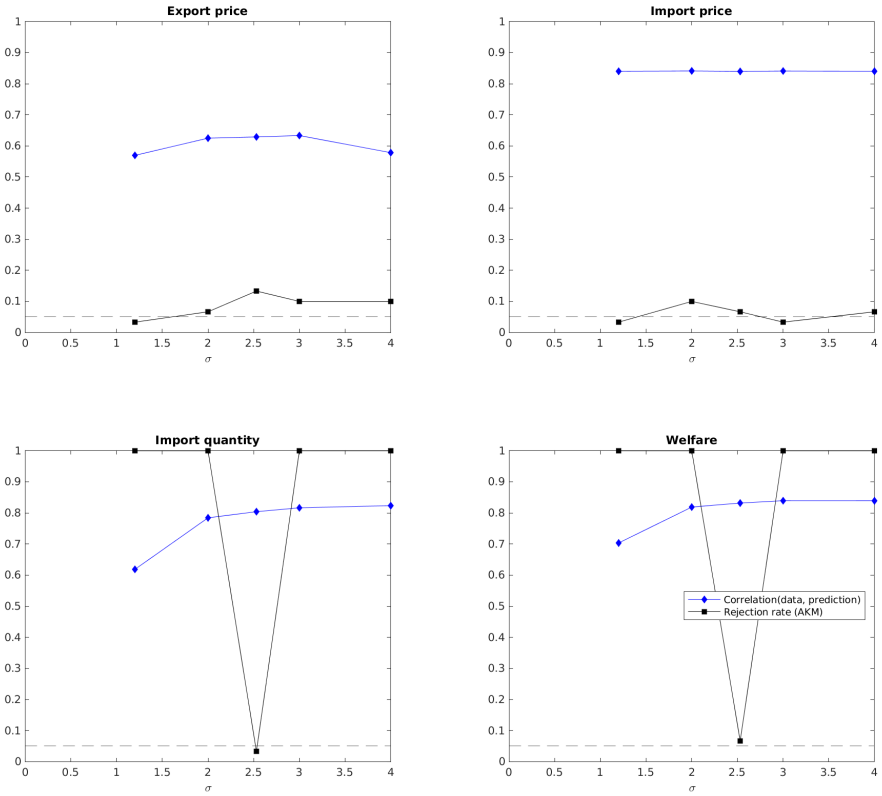
We now turn to an investigation of the power of our test when the researcher's model differs from the true data-generating process. We consider again a sequence of 1,000 hypothetical economies ($b = 1, ..., 1000$) that we calibrate to match the same US data and hit with random shocks drawn from the same distributions used in Section 3.3. We follow the same six-step procedure as before, except that the true model f^* used to generate changes in equilibrium outcomes, $\{\Delta y_n^b = f_n^*(\tau_{t+1}^b, \epsilon_{t+1}^b) - f_n^*(\tau_t^b, \epsilon_t^b)\}$, in Step 5 is now different from that used by the researcher to predict the impact of changes in import tariffs, $\{\Delta x_n^b = f_n(\tau_{t+1}^b, \epsilon_t^b) - f_n(\tau_t^b, \epsilon_t^b)\}$, in Step 6.

[16] This more conservative inference procedure uses as an estimator of the variance of $\beta_{IV} - 1$ given by:

$$\hat{V}_{AKM,0} = \frac{\sum_m \{\sigma \sum_n [(\partial f_n / \partial \tau_m) - \omega_m \sum_r (\partial f_n / \partial \tau_r)][\Delta y_n - \Delta x_n]\}^2}{(\sum_n z_n \Delta x_n)^2}.$$

To do so, we consider a researcher whose model is correct up to the value of the structural elasticities Θ being used. As a first such exercise, we compute the predictions of the researcher's model using the model described in Section 3.1 and the value of the elasticities Θ from Section 3.2, but generate true changes in outcomes using alternative values of import demand elasticities, σ.

Figure 2: Testing when the Researcher's Model is Incorrect (σ)



Notes: This figure reports the rejection rate of the test introduced in Section 2 for Δx_n^b given a 5% significance level (black squares), and the average correlation between Δy_n^b and Δx_n^b (blue diamonds) across $b = 1, ..., 1000$ simulated economies. We generate Δy_n^b and Δx_n^b by calibrating the model to match data for the US in 2017 given the illustrated (on the horizontal axis) value of σ for Δy_n^b and $\sigma = 2.53$ for Δx_n^b. Each panel reports results for the (log of the) outcome indicated in its title, with the bottom-right panel pooling all three outcomes.

Figure 2 reports the rejection rate of our test (black squares) and the average correlation (blue diamonds) across the 1,000 economies that we simulate for each alternative value of σ in the true model, while holding fixed all other elasticities (both for the purposes of generating data and predictions, Δy_n^b and Δx_n^b). Here, the researcher's model maintained that $\sigma = 2.53$. When the true model and the researcher's model agree (that is, at a value of the x-axis corresponding to 2.53) we see that the rejection rate of our test is close to the test's significance level of 5%—just as in the case of Figure 1—since the model

is correct. However, as we consider true values of σ that diverge from 2.53, the rejection rate of our test starts to increase for the import quantity and pooled test cases, but not in the cases of import or export price outcomes. This highlights how the power of a test of a given model can differ dramatically depending on the outcome being tested.[17] On the other hand, the correlation between observed outcomes and the researcher's prediction is not maximized when the researcher's model is correctly specified—a finding that is true for all four outcomes but is particularly apparent in the import quantity and pooled regression cases. That is, while one might hope that the correlation metric for model success would always be strictly monotone in the degree of model misspecification, this hope appears to be misguided.

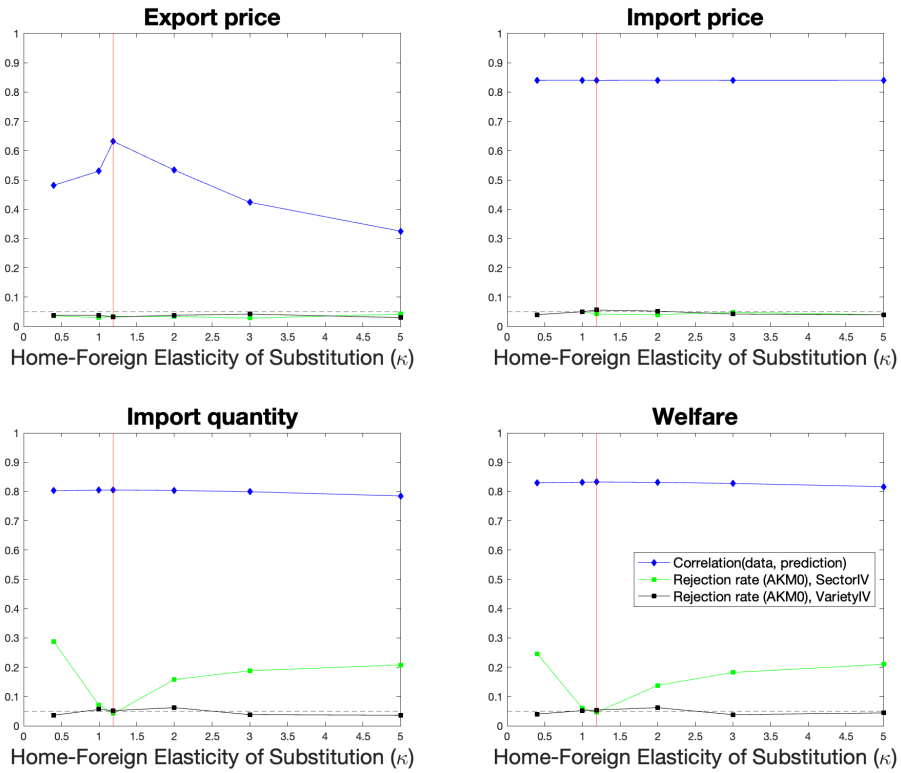
As a second examination of testing in the presence of model misspecification, we now consider a researcher whose model is correct up to the value of the within-sector Home-Foreign elasticity of substitution, κ. As before, Figure 3 reports rejection rates (black squares) and correlations (blue diamonds) for cases in which the researcher's model sets $\kappa = 1.19$, yet the true model takes on a range of alternative values both above and below 1.19. But unlike in the case of σ misspecification presented above, here we see that, for all four outcomes shown, the rejection rate remains close to 5% (even at the upper bound of the range shown, where the true model sets $\kappa = 5$ and the researcher is using $\kappa = 1.19$). That is, our baseline test does not appear to have any power to reject misspecified values of κ in the range displayed.

There is a simple explanation for such a scenario of low-powered tests. The parameter κ governs substitution between the bundle of Home products and that of Foreign varieties (combinations of products and origins), and this substitution therefore occurs at a level of aggregation considerably higher than that of the variety-level observations entering the testing regressions used here. A simple fix is to use an IV that is designed to expose misspecification that turns up at the level of sectors rather than varieties, and a simple form of such an IV is simply the sector-level average of our previous variety-level IV.[18] The green squares in Figure 3 illustrate the rejection rates from our test when using this sector-based IV. As is apparent, the rejection rates show evidence of a test with power against κ-based misspecification, though this is only the case for quantity-based outcomes (import quantities and the pooled regression).

[17] Figure B.1 in Appendix B shows that the higher rejection rates at larger values of $|2.53 - \sigma|$ are a consequence of estimated IV coefficients β that become further from one.

[18] Formally, this is calculated as $z_s \equiv \frac{1}{|\mathcal{I}||\mathcal{G}_s|} \sum_i \sum_{g \in \mathcal{G}_s} z_{ig}$.

Figure 3: Testing when the Researcher's Model is Incorrect (κ)



Notes: This figure reports the rejection rate of the test introduced in Section 2 (with a variety-level IV) for Δx_n^b given a 5% significance level (black squares), the analogous rejection rate for the case of an IV based on sector-level averages of the preveious variety-level IV (green squares), and the the average correlation between Δy_n^b and Δx_n^b (blue diamonds), across $b = 1, ..., 1000$ simulated economies. We generate Δy_n^b and Δx_n^b by calibrating the model to match data for the US in 2017 given the illustrated (on the horizontal axis) value of κ for Δy_n^b and $\kappa = 1.19$ for Δx_n^b. Each panel reports results for the (log of the) outcome indicated in its title, with the bottom-right panel pooling all three outcomes.

4 IV Test Results—using Actual Tariff Changes

Section 3 confirms that the test proposed in Section 2 performs well in simulations of random shocks. We now proceed to ask what it can tell us about the validity of quantitative models in real-world settings. To do so we consider a recent and prominent example of a sudden change in import tariff policy: the Trump Administration's 2018 actions that increased US import tariffs on many products and caused many of its trading partners to retaliate with tariff increases of their own. Fajgelbaum et al.'s (2020) seminal analysis of the 2018 tariff war concluded that "the aggregate real income loss was $7.2 billion, or 0.04% of GDP." We now propose to use our IV test as a way to explore the empirical credibility of this conclusion.

4.1 From Simulated to Actual Shocks

As in the simulations from Section 3, we view FGKK's model as the researcher's model whose predictions we are interested in testing. We therefore again use the model in Section 3.1 as well as the calibrated parameters, Θ and ϵ_t, from Section 3.2, in order to generate predictions Δx_n for the structural responses of the log of import prices, export prices, and import quantities of all varieties. Note that like in Section 3, FGKK's model exactly matches trade and production data on the 2017 US economy, by construction..

We then depart from the simulations in Sections 3.3 and 3.4 in two respects. First, when computing Δx_n, we subject the model economy to actual tariffs in 2017 and 2018 for the US and its trading partners, rather than the randomly drawn tariffs used in our earlier simulations. To do this we average—within each variety, or unique combination of 10-digit HS code products g and foreign countries i, as in Section 3.1—across monthly tariffs within the first four months of 2017 (to obtain $\tau_{ig,t}$ and $\tau^*_{ig,t}$) and the last month of 2018 (for $\tau_{ig,t+1}$ and $\tau^*_{ig,t+1}$), respectively. Second, we use data on actual changes in outcomes Δy_n over 2017-19, rather than simulate those from an hypothetical true model f^*. Under the null that FGKK's model is the true model, this is equivalent to feeding in both the actual tariff shocks (from τ_t to τ_{t+1}) and the non-tariff shocks (from ϵ_t to ϵ_{t+1}) such that FGKK's model also exactly matches data on the 2019 US economy. For each outcome variable of interest, our measure of $y_{n,t}$ corresponds to an average of monthly data for the first four months of 2017, and that for $y_{n,t+1}$ is the equivalent for 2019.

4.2 Evidence from the Trump Tariffs

The results from several versions of our test are reported in Table 1. Each panel reports statistics from a set of IV regressions, following the procedure described in Section 2. We begin with Panel A. Here, the outcome y_n on which the model is being tested is the (log of the) variety-level import price (i.e. $(1+\tau_{ig,t})p^*_{ig,t}$, inclusive of the US import tariff). The sample of observations in Panel A is therefore the 9,289 varieties with a (positive) reported price in both t and $t+1$.

Column 1 reports a specification in which the IV z_n is constructed at the variety level. We see from the reported first-stage F statistic that there is a strong correlation between the model's predicted effect of the tariff changes $\Delta x_n \equiv f_n(\tau_{t+1}, \epsilon_t) - f_n(\tau_t, \epsilon_t)$ and the IV, which is a first-order approximation to this predicted effect. This implies that the model is sufficiently linear, locally, that even the large tariff changes seen in 2018 do not result in a weak IV problem for this regression—and the same is true for essentially all of the specifications in Table 1.

Table 1: IV Test Results

	(1)	(2)	(3)	(4)
Panel A: Log import price				
Coefficient estimate ($\widehat{\beta}_{IV}$)	1.02	0.98	1.56	1.58
standard error	(0.15)	(0.22)	(0.29)	(0.54)
p-value H_0: $\beta_{IV} = 1$	0.91	0.92	0.06	0.28
First-stage F statistic	106,756	69,980	23	76
Observations	9,289	9,289	9,289	9,289
Panel B: Log export price				
Coefficient estimate ($\widehat{\beta}_{IV}$)	-4.56	-2.18	-4.56	-2.18
standard error	(2.28)	(4.54)	(2.28)	(4.54)
p-value H_0: $\beta_{IV} = 1$	0.01	0.48	0.01	0.48
First-stage F statistic	276	289	276	289
Observations	5,860	5,860	5,860	5,860
Panel C: Log import quantity				
Coefficient estimate ($\widehat{\beta}_{IV}$)	0.74	0.54	1.95	-0.03
standard error	(0.12)	(0.27)	(0.49)	(0.63)
p-value H_0: $\beta_{IV} = 1$	0.04	0.09	0.05	0.10
First-stage F statistic	11,407	59,838	9	40
Observations	9,289	9,289	9,289	9,289
Panel D: Pooled welfare outcomes				
Coefficient estimate ($\widehat{\beta}_{IV}$)	0.82	0.98	1.83	1.55
standard error	(0.12)	(0.21)	(0.25)	(0.54)
p-value H_0: $\beta_{IV} = 1$	0.12	0.93	0.00	0.31
First-stage F statistic	23,854	30,790	13	53
Observations	24,438	24,438	24,438	24,438
IV specification:				
Variety-level	Yes	Yes	No	No
Sector-level	No	No	Yes	Yes
Weighted observations	No	Yes	No	Yes

Notes: Each panel and column reports results from a separate IV regression in which each observation is a variety. The dependent variable is the change between 2017 (first four-month average) and 2019 (first four-month average) in the outcome described in the panel heading. The independent variable is the model's prediction for that outcome that would result from the change in tariffs between 2017 (first four-month average) and December 2018. In each case, the IV used is the model's first-order approximation to the effect on the same outcome that would result from the tariff change, starting from a zero-tariff equilibrium. Columns 1 and 2 use an IV based on the model's predicted effects at the variety level and columns 3 and 4 use the sector-level average of such variety-level effects. Weighted regressions use weights: panel A, the variety's 2017 share of import (and in Panel B, export) value in total expenditure; panel C, the 2017 share of tariff revenue in total expenditure; and panel D, each variable as per the weight used in panels A-C.

Turning to the results of the test, in column 1 we estimate $\widehat{\beta}_{IV} = 1.02$. Relative to the the reported standard error (0.15) this estimate is very close to that predicted by the model ($\beta = 1$), so the p-value for the test of the null that $\beta = 1$ is very high (0.91). The IV test, using this outcome, and this instrument, therefore does not reject at standard levels. Column 2 shows that matters are similar when the IV estimator weights each variety-level observation in proportion to that variety's share of total expenditure (in 2017), as would be appropriate for determining (to first-order) the impact of the tariff changes on US consumer welfare via the price of the import bundle.

However, a natural concern when evaluating these results is that the outcome used for testing (the change in each variety's log import price) is closely related to that used for estimation. In particular, all of the within-sector variation in this variable would be guaranteed to lead to a case with $\beta = 1$, so the test's power to reject an invalid specification is unclear. For this reason columns 3 and 4 deploy our IV test with an IV that is based on sector-level averages of the previous variety-level IV. This focuses on the cross-sectoral variation, which was not used for estimation. The estimate of $\widehat{\beta}_{IV}$ is now equal to 1.56 and the p-value is 0.06, implying that this test would reject the model at the 10% level. Finally, column 4 illustrates how a weighted version of this same regression has a very similar point estimate but a larger standard error and hence a higher p-value.

Panel B of Table 1 considers analogous tests for the case when the outcome used is the (log) export price (i.e., $p^X_{ig,t}$, exclusive of any foreign tariff charged). Here, the model's prediction does not vary across varieties within any given sector so the results in columns 1 and 2, which use an IV based on the model's prediction for each variety, are identical to those in columns 3 and 4, which use an IV based on sectoral averages. We see that, whether weighting by the 2017 share of exports in total expenditure (as in column 2) or not (column 1), our estimate of $\widehat{\beta}_{IV}$ is substantially below zero (-2.18 and -4.56, respectively). This results in p-values that reject in the unweighted case (0.01), but do not in the weighted case (0.48). As anticipated by the simulations in Section 3, the model in this setting is inherently harder to test when examining export, in contrast to import price predictions. Reflecting this, the standard errors in Panel B are an order of magnitude larger than those in Panel A.

Turning to Panel C, we next examine the model's predictions for log import *quantities* at the variety level. In this case, the weights used (in columns 2 and 4) are based on each variety's 2017 share of tariff revenue in total expenditure, as motivated by the first-order approximation to the fiscal externality effect of the tariff changes on welfare. Here, columns 1 and 2 again use a variety-level based IV, which contains a mixture of within-sector and between-sector variation. As such, there are mechanical forces pushing the

coefficients towards one. However, when we focus on the sector-averaged IV (in columns 3 and 4) this mechanical component is removed, and this tends to push the coefficients away from one (to 1.95 and -0.03, respectively) as well as result in model rejection at the 10% level.

The final test that we consider (as reported in Panel D of Table 1) aims to provide an omnibus test of the model's overall predictions and gets us as close as possible, given available IVs, to the welfare question of interest. To do so, we pool each of the three aforementioned outcomes (import prices, export prices, and import quantities) into one larger regression so that the testing coefficient β is an assessment of the model's ability to accurately predict the effect of tariff changes on any one of these observations. The weighted regressions in columns 2 and 4 now take on a particular welfare-oriented interpretation, since each observation is weighted in a way that is suggestive of its role in the contribution of a price or quantity change to aggregate welfare. Given the low level of tariffs in 2017, this means that the import quantity effects (weighted in proportion to a variety's contribution to tariff revenue) will matter substantially less than those for import and export prices (whose weight is proportional to trade values).

Focusing on these weighted regressions, we see (in column 2) that the variety-level IV specification yields an estimate ($\widehat{\beta}_{IV} = 0.98$) that is very close to one. The model performs less well ($\widehat{\beta}_{IV}$ rises to 1.55) when tested via the sector-level IV—which again avoids using mechanical fit when testing—in column 4. This point estimate implies that the model's weighted predictions are off by more than a factor of $1/3$ but the test for the null of $\beta = 1$ does not reject at standard levels ($p = 0.31$).

We therefore view the results in Panel D as offering credibility to the model's answer to the counterfactual question being posed of it. This contrasts with the impression one could get by examining a simple correlation between the model's prediction Δx_n and what actually happened in the data Δy_n. The weighted version of this correlation, at the variety level, for the pooled set of outcomes is approximately 0.1—corresponding to an R^2 of just 0.011—and that for every other outcome in Table 1, whether weighted or not, is even lower. Ultimately, the variance of shocks to ϵ_t in this setting appears to be substantially higher than that of tariff changes, but the relative size of these two variances is inconsequential for our IV test.

5 Concluding Remarks

An important goal of a large trade literature is to quantify how a country's economy would respond if a given change in tariffs were to be implemented. To help assess and po-

tentially strengthen the credibility of such quantitative trade model predictions we have proposed a new testing procedure that is intuitive and easy to implement. Critically, unlike standard goodness of fit tests, our test is applicable even in the presence of other unobserved shocks that may affect the economy of interest, and also in settings where the model is sufficiently flexible that it can match all observations. Our test exploits policy changes, or observable shifters of these, that are orthogonal to unobserved shocks, a standard exclusion restriction consistent with existing estimation procedures in the field. As an illustration of how our procedure can be combined with state-of-the-art quantitative work, we have tested the welfare consequences of the Trump tariffs predicted by Fajgelbaum et al. (2020). The preferred version of our test shows that the model cannot be rejected at standard levels, thereby offering credibility to the model's answer to the counterfactual question being posed of it.

References

Adao, Rodrigo, Arnaud Costinot, and Dave Donaldson, "Nonparametric Counterfactual Predictions in Neoclassical Models of International Trade," *American Economic Review*, 2017, *107* (3), 633–689.

_ , Costas Arkolakis, and Federico Esposito, "General equilibrium effects in space: Theory and measurement," 2020. Working paper.

_ , Michal Kolesar, and Eduardo Morales, "Shift-Share Designs: Theory and Inference," *The Quarterly Journal of Economics*, 2019, *134* (4), 1949–2010.

_ , Paul Carillo, Arnaud Costinot, Dave Donaldson, and Dina Pomeranz, "Imports, Exports, and Earnings Inequality: Measures of Exposure and Estimates of Incidence," *Quarterly Journal of Economics*, forthcoming.

Attanasio, Orazio, Pinelopi Goldberg, and Nina Pavcnik, "Trade Reforms and Wage Inequality in Colombia," *Journal of Development Economics*, 2004, *74* (2), 331–366.

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

Borusyak, Kirill, Peter Hull, and Xavier Jaravel, "Quasi-experimental shift-share research designs," *Review of Economic Studies*, 2022, *89* (1), 181–213. Forthcoming.

Bowen, Harry P., Edward E. Leamer, and Leo Sveikauskas, "Multicountry, Multifactor Tests of the Factor Abundance Theory," *American Economic Review*, 1987, *77* (5), 791–809.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro, "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock," *Econometrica*, 2019, *87* (3), 741–835.

Costinot, Arnaud and Andres Rodríguez-Clare, "Trade Theory with Numbers: Quantifying the Consequences of Globalization," in Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds., *Handbook of International Economics*, Vol. 4, New York: Elsevier, 2014.

_ and Dave Donaldson, "Ricardo's Theory of Comparative Advantage: Old Idea, New Evidence," *American Economic Review, Papers and Proceedings*, 2012, *102* (3), 453–458.

Davis, Donald R. and David Weinstein, "An Account of Global Factor Trade," *American Economic Review*, 2001, *91* (5), 1423–1453.

Dawkins, Christina, T.N. Srinivasan, and John Whalley, *Calibration*, Vol. 5 of *Handbook of Econometrics*, Elsevier, 2001.

Dingel, Jonathan and Felix Tintelnot, "Spatial Economics for Granular Settings," *BFI Working Paper 2020-71*, 2021.

Eaton, Jonathan, Samuel Kortum, Brent Neiman, and John Romalis, "Trade and the Global Recession," *American Economic Review*, 2016, *106* (11), 3401–3438.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J Kennedy, and Amit K Khandelwal, "The Return to Protectionism," *Quarterly Journal of Economics*, 2020, *135* (1), 1–55.

Goldberg, Pinelopi K. and Nina Pavcnik, *The Effects of Trade Policy*, Vol. 1A of *Handbook of Commercial Policy*, Elsevier, 2016.

Kehoe, Timothy J., "An Evaluation of the Performance of Applied General Equilibrium Models of the Impact of NAFTA," in Timothy J. Kehoe, T.N. Srinivasan, and John Whalley, eds., *Frontiers in Applied General Equilibrium Modeling*, New York: Cambridge University Press, 2005, pp. 341–377.

_ , Clemente Polo, and Ferran Sancho, "An evaluation of the performance of an applied general equilibrium model of the spanish economy," *Economic Theory*, 1995, pp. 115–141.

_ , Pau S. Pujolas, and Jack Rossbach, "Quantitative Trade Models: Developments and Challenges," *Annual Review of Economics*, 2017, *9*, 295–325.

Kovak, Brian K., "Regional Effects of Trade Reform: What Is the Correct Measure of Liberalization?," *American Economic Review*, 2013, *103* (5), 1960–76.

Lai, Huiwen and Daniel Trefler, "The Gains from Trade with Monopolistic Competition: Specification, Estimation, and Mis-Specification," *NBER Working Paper*, 2002.

Leamer, Edward E. and James E. Levinsohn, *International Trade Theory: The Evidence*, Vol. 3 of *Handbook of International Economics*, Elsevier, 1995.

McCaig, Brian, "Exporting out of poverty: provincial poverty in Vietnam and U.S. market access," *Journal of Intermational Economics*, 2011, *85* (1), 102–113.

Redding, Stephen and Esteban Rossi-Hansberg, "Quantitative Spatial Economics," *Annual Review of Economics*, 2017, *9*, 21–58.

Shoven, John B. and John Whalley, "Applied General-Equilibrium Models of Taxation and International Trade: An Introduction and Survey," *Journal of Economic Literature*, 1984, *22* (3), 1007–1051.

Topalova, Petia, "Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India," *American Economic Journal: Applied Economics*, 2010, *2* (4), 1–41.

Trefler, Daniel, "The Case of the Missing Trade and Other Mysteries," *American Economic Review*, 1995, *85* (5), 1029–46.

A Theoretical Appendix

A.1 Consistency of β_{IV}

Consider the instrument variable and structural residuals,$z_n \equiv \sum_m (\partial f_n/\partial \tau_m)(\Delta\tau_{IV,m} - \Delta\bar{\tau}_{IV})$ as our IV, where $\Delta\bar{\tau}_{IV}$ is a weighted average of the initial tariffs, $\Delta\bar{\tau}_{IV} \equiv \frac{1}{M}\sum_m \omega_m \Delta\tau_{IV,m}$

$$z_n = \sum_m (\partial f_n/\partial \tau_m)(\Delta\tau_{IV,m} - \Delta\bar{\tau}_{IV}), \tag{A.1}$$

$$\Delta\eta_n^* = f_n(\tau_{t+1}, \epsilon_{t+1}) - f_n(\tau_{t+1}, \epsilon_t), \tag{A.2}$$

where $\Delta\bar{\tau}_{IV} \equiv \frac{1}{M}\sum_m \omega_m \Delta\tau_{IV,m}$ denotes a weighted average of tariff shifters, for some fixed weights $\sum_m \omega_m = 1$, and the gradient $\{\partial f_n/\partial\tau_m\}$ is evaluated at some fixed value of the vector of tariffs and other shocks, (τ, ϵ).

The goal of this appendix is to describe sufficient conditions on the data generating process for $(\Delta\tau_{IV}, \tau_{t+1}, \epsilon_t, \epsilon_{t+1})$ such that the exclusion restriction, $\text{plim}_{N\to\infty}\frac{1}{N}\sum_n z_n \Delta\eta_n^* = 0$, invoked in Proposition 1 to establish the consistency of β_{IV} holds. We do so through a series of three lemmas that follow closely the arguments in Borusyak et al. (2022).

Lemma 1. *Suppose that* (i) $E[\frac{1}{N}\sum_n z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = 0$ *and* (ii) $Var[\frac{1}{N}\sum_n z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \to 0$. *Then* $plim_{N\to\infty}\frac{1}{N}\sum_n z_n \Delta\eta_n^* = 0$.

Proof. Condition (i) implies $\text{Var}[\frac{1}{N}\sum_n z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = E[(\frac{1}{N}\sum_n z_n \Delta\eta_n^*)^2 | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}]$. Thus we can rearrange condition (ii) as $E[(\frac{1}{N}\sum_n z_n \Delta\eta_n^*)^2 | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \to 0$. Since convergence in mean square implies convergence in probability, It follows that $\frac{1}{N}\sum_n z_n \Delta\eta_n^* \to_p 0$. □

Lemma 2. *Suppose that* (i) $\Delta\tau_{IV,m}$ *are i.i.d across* m *and* (ii) $\Delta\tau_{IV,m}$ *are independent of* $(\tau_{t+1}, \epsilon_t, \epsilon_{t+1})$. *Then* $E[\frac{1}{N}\sum_n z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = 0$.

Proof. Start from

$$E[\frac{1}{N}\sum_n z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = \frac{1}{N}\sum_n E[z_n \Delta\eta_n^* | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = \frac{1}{N}\sum_n \Delta\eta_n^* E[z_n | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}], \tag{A.3}$$

where the second equality uses the definition of $\Delta\eta_n^*$ in equation (A.2). By condition (i) and the definition of z_n in equation (A.1), we also have

$$\begin{aligned} E[z_n | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] &= E[\sum_m (\partial f_n/\partial\tau_m)(\Delta\tau_{IV,m} - \Delta\bar{\tau}_{IV}) | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \\ &= \sum_m (\partial f_n/\partial\tau_m)(E[\Delta\tau_{IV,m} | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}] - \sum_r \omega_r E[\Delta\tau_{IV,r} | \tau_{t+1}, \epsilon_t, \epsilon_{t+1}]) = 0. \end{aligned} \tag{A.4}$$

Lemma 2 follows from equations (A.3) and (A.4). □

Lemma 3. *Suppose that* (i) $\Delta\tau_{IV,m}$ *are i.i.d across* m; (ii) $\Delta\tau_{IV,m}$ *are independent of* $(\tau_{t+1}, \epsilon_t, \epsilon_{t+1})$; (iii) $Var[\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = \sigma^2 < \infty$; (iv) $\Delta\eta_n^* \leq M$ *for all* n; *and* (v) $\frac{1}{N^2}\sum_m(\alpha_{m,t})^2 \to 0$, *with* $\alpha_{m,t} \equiv \sum_n[(\partial f_n/\partial\tau_m) - \omega_m \frac{1}{N}\sum_r(\partial f_n/\partial\tau_r)]$. *Then* $Var[\frac{1}{N}\sum_n z_n\Delta\eta_n^*|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \to 0$.

Proof. Start from

$$\begin{aligned}
&\text{Var}[\frac{1}{N}\sum_n z_n\Delta\eta_n^*|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \\
&\qquad = \text{Var}[\frac{1}{N}\sum_m[\sum_n \Delta\eta_n^*[(\partial f_n/\partial\tau_m) - \omega_m\sum_r(\partial f_n/\partial\tau_r)]\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \\
&\qquad = \frac{1}{N^2}\text{Var}[\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \\
&\qquad\qquad \times \sum_m(\sum_n \Delta\eta_n^*[(\partial f_n/\partial\tau_m) - \omega_m\sum_r(\partial f_n/\partial\tau_r)]\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}])^2.
\end{aligned}$$

where the first equality derives from the definition of z_n in equation (A.1) and the second equality from conditions (i) and (ii). Using conditions (iii) and (iv) and the definition of $\alpha_{m,t} \equiv \sum_n[(\partial f_n/\partial\tau_m) - \omega_m\sum_r(\partial f_n/\partial\tau_r)]$, we further get

$$\text{Var}[\frac{1}{N}\sum_n z_n\Delta\eta_n^*|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] \leq \frac{\sigma^2M^2}{N^2}\sum_m(\alpha_m)^2.$$

Lemma 3 follows from the previous inequality and condition (v). □

Combining the three previous lemmas, we obtain the following sufficient conditions on the data generating process for tariffs and other shocks $(\tau_t, \tau_{t+1}, \epsilon_t, \epsilon_{t+1})$ such that the exclusion restriction invoked in Proposition 1, $\text{plim}_{N\to\infty}\frac{1}{N}\sum_n z_n\Delta\eta_n^* = 0$, holds.

Proposition 3. *Suppose that* (i) $\Delta\tau_{IV,m}$ *are i.i.d across* m *and* (ii) $\Delta\tau_{IV,m}$ *are independent of* $(\tau_{t+1}, \epsilon_t, \epsilon_{t+1})$; (iii) $Var[\Delta\tau_{IV,m}|\tau_{t+1}, \epsilon_t, \epsilon_{t+1}] = \sigma^2 < \infty$; (iv) $\Delta\eta_n^* \leq M$ *for all* n; *and* (v) $\frac{1}{N^2}\sum_m(\alpha_{m,t})^2 \to 0$, *with* $\alpha_{m,t} \equiv \sum_n[(\partial f_n/\partial\tau_m) - \omega_m\sum_r(\partial f_n/\partial\tau_r)]$. *Then* $plim_{N\to\infty}\frac{1}{N}\sum_n z_n\Delta\eta_n^* = 0$.

A.2 Asymptotic Distribution of β_{IV}

The goal of this appendix is to provide the proof of Proposition 2 following the general argument in Adao et al. (2019).

Proof of Proposition 2. By definition of β_{IV}, we have

$$\frac{\sum_n z_n\Delta x_n}{(\sum_m(\alpha_{m,t})^2)^{1/2}}(\beta_{IV} - 1) = \frac{\sum_n z_n\Delta x_n}{(\sum_m(\alpha_{m,t})^2)^{1/2}}\left(\frac{\frac{1}{N}\sum_n z_n\Delta y_n}{\frac{1}{N}\sum_n z_n\Delta x_n} - 1\right) = \frac{\sum_n z_n\Delta\eta_n^*}{(\sum_m(\alpha_{m,t})^2)^{1/2}}.$$

Let $R_m \equiv \sum_n [(\partial f_n/\partial \tau_m) - \omega_m \sum_r (\partial f_n/\partial \tau_r)]\Delta\eta_n^*$. Using the previous notation, we can write

$$\frac{\sum_n z_n \Delta\eta_n^*}{\left(\sum_m (\alpha_{m,t})^2\right)^{1/2}} = \frac{\sum_m R_{m,t}\Delta\tau_{IV,m}}{\left(\sum_m (\alpha_{m,t})^2\right)^{1/2}}.$$

Next, consider $\mathcal{Y}_m \equiv R_m \Delta\tau_{IV,m}$ and the filtration $\mathcal{F}_m \equiv \sigma(\Delta\tau_{IV,1}, ..., \Delta\tau_{IV,m}, \tau_{t+1}, \epsilon_t, \epsilon_{t+1})$. Since $\Delta\tau_{IV,m}$ are i.i.d across m and independent of $(\tau_{t+1}, \epsilon_t, \epsilon_{t+1})$, $E[\mathcal{Y}_m|\mathcal{F}_{m-1}] = 0$ for all m. Thus, $\mathcal{Y}_m$ is a martingale difference array with respect to the filtration $\mathcal{F}_m$. It satisfies

$$\begin{aligned}
\frac{\sum_m E\left[\mathcal{Y}_m^4|\tau_{t+1},\epsilon_t,\epsilon_{t+1}\right]}{[\sum_m (\alpha_{m,t})^2]^2} &= \frac{\sum_m R_{m,t}^4 E\left[(\Delta\tau_{IV,m})^4|\tau_{t+1},\epsilon_t,\epsilon_{t+1}\right]}{[\sum_m (\alpha_{m,t})^2]^2} \\
&\leq T\frac{\sum_m R_{m,t}^4}{[\sum_m (\alpha_{m,t})^2]^2} \\
&= T\frac{\sum_m [\sum_n \delta_{nm}\Delta\eta_n^*]^4}{[\sum_m (\alpha_{m,t})^2]^2} \\
&= T\frac{\sum_m \left(\sum_i \sum_j \delta_{jm}\delta_{im}\Delta\eta_i^*\Delta\eta_j^*\right)^2}{[\sum_m (\alpha_{m,t})^2]^2} \\
&\leq TM^2\frac{\sum_m \left(\sum_i \sum_j \delta_{jm}\delta_{im}\right)^2}{[\sum_m (\alpha_{m,t})^2]^2} \\
&= TM^2\frac{\sum_m (\alpha_{m,t})^4}{[\sum_m (\alpha_{m,t})^2]^2} \leq TM^2\frac{\left(\max_m (\alpha_{m,t})^2\right)}{\sum_m (\alpha_{m,t})^2},
\end{aligned}$$

where the first inequality derives from $E[(\tau_{m,t})^4|\tau_{t+1},\epsilon_t,\epsilon_{t+1}] \leq T$, the second inequality derives from $\Delta\eta_n^* \leq M$ for all n, and $\delta_{nm} \equiv (\partial f_n/\partial\tau_m) - \omega_m \sum_r (\partial f_n/\partial\tau_r)$. Combining the final inequality with condition $\max_m(\alpha_{m,t})/\sum_r \alpha_{r,t}^2 \to 0$, we get $\sum_m E\left[\mathcal{Y}_m^4|\tau_{t+1},\epsilon_t,\epsilon_{t+1}\right][\sum_m (\alpha_{m,t})^2]^{-2} \to 0$. Hence there exists $v > 0$ such that $[\sum_m (\alpha_{m,t})^2]^{-(1+v/4)}\sum_m E\left[\mathcal{Y}_m^{2+v/2}|\tau_{t+1},\epsilon_t,\epsilon_{t+1}\right] \to 0$. Since all the assumptions of Proposition 3 are satisfied, we also know that $\text{plim}_{N\to\infty}\frac{1}{N}\sum_m \mathcal{Y}_m = 0$. We can therefore invoke the dominated convergence theorem and martingale central limit theorem to establish that

$$\frac{\sum_n z_n\Delta\eta_n^*}{\left(\sum_m (\alpha_{m,t})^2\right)^{1/2}} = \frac{\sum_m \mathcal{Y}_m}{\left(\sum_m (\alpha_{m,t})^2\right)^{1/2}} \to_d \mathcal{N}(0, V_N),$$

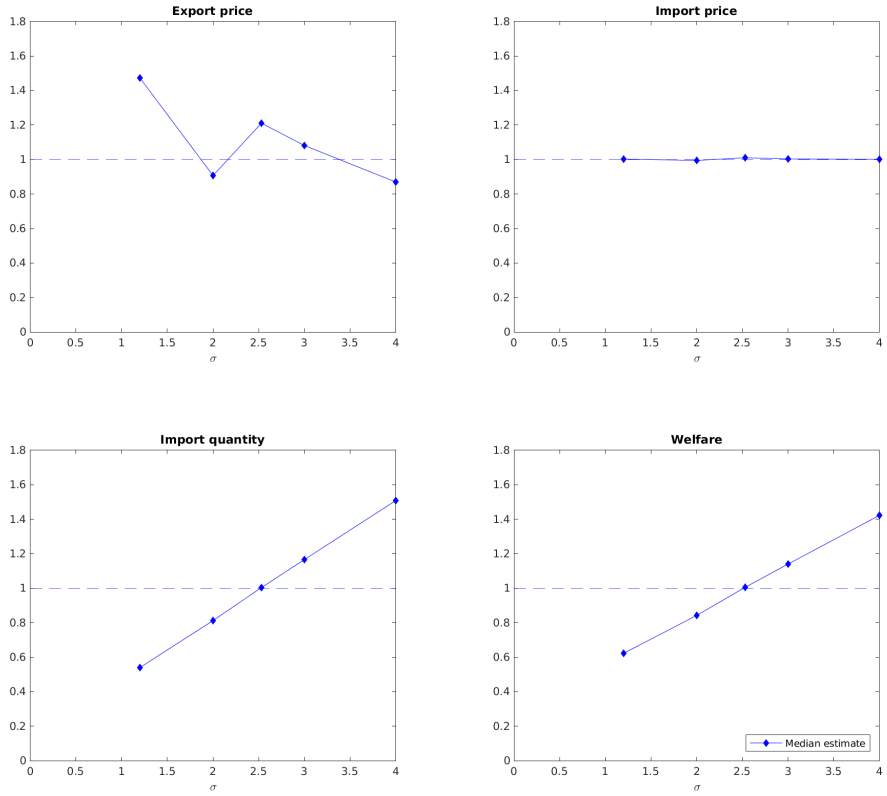
where the variance V_N is such that

$$\begin{aligned}
V_N = \frac{\sum_m E\left[\mathcal{Y}_m^2|\mathcal{F}_{m-1}\right]}{\sum_m (\alpha_{m,t})^2} &= \frac{\sum_m E\left[(\Delta\tau_{IV,m})^2 R_{m,t}^2|\mathcal{F}_{m-1}\right]}{\sum_m (\alpha_{m,t})^2} \\
&= \frac{\sum_m E\left[(\Delta\tau_{IV,m})^2|\mathcal{F}_{m-1}\right]R_{m,t}^2}{\sum_m (\alpha_{m,t})^2} = \frac{\sum_m \sigma^2 R_{m,t}^2}{\sum_m (\alpha_{m,t})^2}.
\end{aligned}$$

□

B Test Results Appendix

Figure B.1: Median estimated coefficient of testing specification, when model calibration of σ is incorrect



Notes: The figure reports the median estimated coefficient of our testing specification across $b = 1, ..., 1000$ simulated economies. We generate Δy_n^b and Δx_n^b by calibrating the model to match data for the US in 2017 given the indicated (on the horizontal axis) value for σ for Δy_n^b and $\sigma = 2.53$ for Δx_n^b. Each panel reports results for the outcome indicated in its title, with the bottom-right panel pooling all three outcomes.