CHAPTER 18

Trade Elasticity Parameters for a Computable General Equilibrium Model

Russell Hillberry*, David Hummels**
*Department of Economics, University of Melbourne
**Purdue University, and National Bureau of Economic Research

Abstract

Computable general equilibrium (CGE) models of international trade typically rely on econometrically estimated trade elasticities as model inputs. These elasticities vary by as much as an order of magnitude and there is no consensus on which elasticities to use. We review the literature estimating trade elasticities, focusing on several key considerations. What are the identifying assumptions used to separate supply and demand parameters? What is the nature of the shock to prices employed in the econometrics? And what is the time horizon over which trade responds to this shock? This discussion ranges from older reduced form approaches that use time-series variation in prices, to more recent work that identifies demand elasticities from trade costs or by using instruments in cross-section or panel data, and finally to prominent applications that separately identify supply and demand parameters in the absence of instruments. We also discuss recent theoretical developments from the literature on heterogeneous firms that complicate the interpretation of all these parameter estimates. Finally, we briefly survey a literature on structural estimation and link this to recent attempts to incorporate such theories in CGE applications. By elucidating the differences and similarities in these approaches we hope to guide the CGE practitioner in choosing elasticity estimates. We favor elasticities taken from econometric exercises that employ identifying assumptions and exploit shocks that are similar in nature to those imposed in the model experiment.

Keywords

Import demand elasticity, export supply elasticity, elasticity of substitution

JEL classification codes

F00, F17, F19

18.1 INTRODUCTION

This chapter discusses trade elasticities — the response of traded quantities to changes in prices of tradable goods. While the results of computable general equilibrium (CGE) experiments depend upon a number of inputs, trade elasticities are of particular interest

Handbook of CGE Modeling - Vol. 1 SET
ISSN 2211-6885, http://dx.doi.org/10.1016/B978-0-444-59568-3.00018-3

because they significantly impact upon the modeled effects of policy experiments on trade patterns, welfare and factor returns, among other important phenomena.

It is common when calibrating CGE models to select trade elasticities from "the literature" while selecting other (taste and technology) parameters to allow the theory to replicate the data.[1] Curiously, there is no clear consensus on which elasticities to use. Major trade-focused CGE models draw elasticities from many different econometric studies. These econometric studies use very different data samples, response horizons and estimating techniques, and arrive at elasticities as much as an order of magnitude different from each other. This raises the critical question: which elasticities are "right?" Or at least, which are right for the particular modeling application at hand?

As a starting point for thinking about these issues, Figure 18.1 presents a simple partial equilibrium diagram in which the price and quantity of traded goods depends on export supply and import demand. Using this diagram, we can think through the effects of a policy experiment such as raising a tariff on foreign goods.

To fix ideas, consider a parsimonious representation of import demand in which quantities imported depend on prices in the foreign country (F), inclusive of tariffs and real expenditures in the home country (H):

$$\ln q_{\mathrm{F}} = \ln E_{\mathrm{H}} - \sigma \ln p_{\mathrm{F}}(1+\tau). \tag{18.1}$$

The key parameter is σ, which can be thought of as a reduced form measuring the elasticity of import quantities with respect to import prices, but is more commonly given a structural interpretation. For example, in many common CGE frameworks, this demand function arises from a constant elasticity of substitution (CES) cost or utility function in which buyers regard home and foreign varieties as imperfect substitutes. This is known as the Armington assumption (and σ is sometimes referred to as the Armington

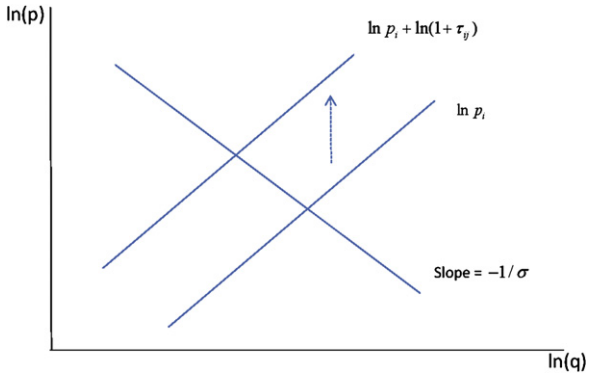

Figure 18.1 Import demand and export supply with a tariff shock.

[1] This approach can be attributed to Shoven and Whalley (1972), although the use of externally estimated parameters from the econometric literature came later.

parameter or Armington elasticity), although very similar formulations arise in other common modeling frameworks such as monopolistic competition.

The parameterization of σ has important quantitative implications for a number of variables that are of interest to economists and policy makers alike, and we highlight these in Section 18.2. The most direct and explicit link is via import quantities. In Figure 18.1, a rise in tariff rates shifts the export supply curve upwards along the import demand curve. Here, the elasticity of import demand effectively summarizes the first-order response of traded quantities to changes in trade cost changes. These first-order effects, as summarized in Equation (18.2) imply that doubling the trade elasticity will double the response in measured quantities.[2]

In Section 18.3 we survey the literature estimating import demand elasticities. We highlight important differences across the econometric literature in the price shocks observed, the time horizon over which responses are measured, the comparison set of countries and the level of aggregation. Estimates of σ vary considerably, and we provide a lengthy discussion of why these estimates vary and which are appropriate in different circumstances.

A recurring theme throughout the chapter is the difficulty of separating supply and demand parameters. Ideally, one would observe movements in export supply induced by policy shocks in the manner described in Figure 18.1. Unfortunately, data experiments of this sort are somewhat rare and so many early studies exploit time-series variation in foreign prices p_F in Equation (18.1). Since prices are jointly determined by supply and demand this raises a critical issue of identification: are these time-series studies observing shocks to supply and identifying the elasticity of import demand or are they observing shocks to demand and observing the elasticity of export supply, or some combination of the two? In more recent econometric papers we survey, price variation is driven by shocks to tariffs or transportation costs in precisely the manner described in Figure 18.1. This sort of estimation procedure provides a reasonably close match to thought experiments typically contemplated in CGE trade liberalization exercises and also allows the econometrician to better control for shocks to demand.

In Section 18.4 we turn to the literature on estimating the elasticity of export supply. Single-country CGE trade models do not provide an explicit modeling of production and demand in the rest of the world. Instead, they may parameterize a country's exports to the rest of the world (and the supply of imports into that country from the rest of the world) in a reduced form way. These approaches have a weak connection between the underlying supply-side details that give rise to an

[2] Shocks to delivered prices (often induced by changes in tariff rates) generate first-order changes in the relative demand for two varieties. If general equilibrium responses (in prices) to this change are sufficiently small, then the structural demand parameter σ is the relevant trade elasticity in the sense that it largely describes the total trade response to a trade policy shock. That is to say, a tariff cut that lowers the price of foreign relative to home goods by 1% results in a σ% increase in imports.

export supply curve as in Figure 18.1 and we discuss the reduced form econometric work used to parameterize it.

Multicountry CGE trade models do provide explicit modeling of production and demand worldwide, and so do not parameterize export supply in this reduced form way. While these models are primarily interested in import demand elasticities, the identification issue just discussed requires the econometrician to account for supply. We next discuss a literature that estimates systems of export supply and import demand in order to get proper identification of each. While this literature has primarily been mined for import demand elasticities, it provides a potentially useful source of export supply elasticities.

To make progress on the export supply front it is necessary to move away from reduced forms and provide a parameterization that is closely linked to theory. We discuss developments in the literature on trade with heterogeneous firms that provide such a link. These developments are useful on one dimension and challenging on another — the possibility of within industry heterogeneity calls into question the identification of demand parameters used throughout a large literature, they suggest that econometricians are actually, and only, estimating supply responses!

Finally, in Section 18.5 we discuss structural estimation as a possible way forward. We step the reader through a progression of papers in order to show the assumptions under which import demand and export supply parameters can be extracted from available data. None of these approaches are "magic bullets." Our discussion instead highlights the point that these papers differ primarily in what they hold fixed, or what external parameters they bring to bear in order to extract residual parameters from the data.

Ultimately, the interpretation of trade responses is model dependent. CGE practitioners should come away from this survey with a sense of where the elasticities "in the literature" come from, how they are identified and what they purport to measure. We do not ultimately pronounce upon the question of which estimates provide the "right" elasticities. Rather we hope to inform the choice of trade response parameters by informing practitioners about the nature of the assumptions econometricians have undertaken in order to move from the theory to the data to resulting estimates.

18.2 WHY DO TRADE ELASTICITIES MATTER?

Consider a simple representation of CES utility and its associated relative demand:

$$U = \left(\sum_c b_c q_c^{\theta}\right)^{1/\theta}, \quad \theta = (\sigma - 1)/\sigma$$

$$\ln\frac{q_i}{q_j} = \frac{b_i}{b_j} - \sigma \ln\left(\frac{p_i(1+\tau_i)}{p_j(1+\tau_j)}\right). \tag{18.2}$$

The summation in the utility function runs over distinct product varieties. Here, we employ the Armington assumption that buyers treat varieties as differentiated on the basis of country of origin (indexed by c). Relative demand for goods originating in source countries i and j depends on relative prices inclusive of *ad valorem* tariffs $p(1+\tau)$, and some additional terms b that capture "tastes" or other non-price supply factors such as quality or variety.

It is no exaggeration to say that σ is the most important parameter in modern trade theory. σ captures both the own-price elasticity of demand and the (inverse) cross-price elasticity of demand, and it is the elasticity of substitution between two varieties measuring how "close" the two goods are in product space. As such, σ is critical for evaluating welfare gains from price changes and the provision of new variety as in Feenstra (1994) and Broda and Weinstein (2006). In monopolistic competition models with this preference structure such as Krugman (1980) and the many papers built on it, σ governs both the scale of the firm and markups over marginal cost.[3] In models of economic geography following Krugman (1991), σ governs the strength of agglomeration economies and home market effects. In empirical work on border costs and distance effects, and the literature on explaining rapid growth in trade, σ is of central importance in determining the level and changes in trade costs needed to match facts on trade volumes. In models of heterogeneous firms and trade following Melitz (2003), σ partly determines the size distribution of firms and participation in export markets.

Not surprisingly then, trade elasticities are of central importance for quantitative analysis of trade policy. The most direct and explicit link, as discussed in the introduction, is between price or tariff changes and the associated change in traded quantities. For example, trade elasticities play a central role in determining the effects of preferential trade agreements. As noted as early as Viner, the welfare consequences of a preferential agreement depend on whether the agreement creates additional trade between the parties, or simply diverts trade away from partners outside the agreement. When incorporated in CGE models, the strength of these trade creation and trade diversion effects are largely dictated by the Armington elasticity.[4]

As trade elasticities play a large role in determining the size and nature of trade adjustment to policy shocks, they also impact other modeled economic phenomena that are affected by changes in trade volumes. Many trade theories suggest that trade policy changes have important distributional effects within national economies. In models with

[3] In this instance, we assume that each firm produces a unique variety. If firms within a country are homogeneous we can still sum over countries by first replacing taste parameters b with the number of firms producing in each country.

[4] The US International Trade Commission (2004), in its assessment of the US–Australia Free Trade Agreement, undertook a systematic sensitivity analysis of model results, with the source of the underlying uncertainty being the statistical uncertainty attached to the estimates in an econometric exercise conducted by Hertel *et al.* (2007). The point estimate of the US gain (in equivalent variation) attached to the agreement was $491 million, while the 95% confidence interval ranged from $435 to $639 million.

a role for trade policy-induced productivity growth, productivity growth depends on the parameterization of the trade response.[5] Perhaps the most salient are those cases in which the effectiveness of specific policies are determined in part by the trade environment. This can be true for a wide variety of policies, including developed country agricultural support, tax policy and environmental policy.

A central variable of interest is welfare. The link between trade elasticities and welfare has been explored at length in recent theory surrounding the gravity model of trade. Arkolakis *et al.* (2010) show that a wide variety of trade models link trade elasticities and welfare in the same manner. In these models, the percentage change in real income $\hat{W}$ can be summarized by the percentage change in the share of domestic expenditure on domestic output ($\hat{\lambda}$) and the elasticity of trade (value) to trade price changes, ε:[6]

$$\hat{W} = \hat{\lambda}^{1/\varepsilon}. \tag{18.3}$$

The broad class of models assessed by Arkolakis *et al.* remain special cases, as they rely on particular representations of trade costs, factor supplies and preferences that are usually not replicated in CGE studies. It is clear nonetheless that trade elasticities play a central role in CGE exercises. Valenzuela *et al.* (2008), for example, conduct sensitivity analysis over alternate parameterizations of Armington substitution elasticities in an assessment of the likely impacts of trade liberalization in the Doha round. They find that doubling the Armington elasticities roughly doubles both the trade response and the welfare gains in the Global Trade Analysis Project (GTAP) model.

Changes in the terms of trade are one important channel through which the elasticity might matter for welfare. Brown (1987) provides a provocative study linking the choice of σ to the ability of even small countries to affect the terms of trade through optimal tariffs. When modelers employ Armington preferences they are treating each country as producing a unique variety of each good. This means that even small countries have market power they can use to affect their terms of trade. Just how much market power they have depends on whether buyers view national varieties as highly substitutable or very distinct. For small values of σ, all countries enjoy substantial terms-of-trade power. In this case, optimal tariffs often lie above observed applied rates, which implies that unilateral trade liberalization is likely to be welfare reducing.

This implication sits uncomfortably with textbook treatments of the issue, in which small countries are thought to have optimal tariffs at or near zero. Moreover, the policy implication is difficult to present to policy makers, as it contradicts mainstream trade

[5] This link is most clearly made in recent models of heterogeneous firms, such as Melitz (2003). In these models, the nature of the trade response is that high productivity export firms grow while low productivity firms exit. The extent of this reallocation (and the associated productivity gain associated with trade liberalization) is governed by the single parameter that defines the trade response and the shape of the firm productivity distribution.

[6] The notation ε is general, representing any trade elasticity. In the context of an Armington model like that suggested by, $\varepsilon = 1 - \sigma$.

theory on the merits of unilateral trade liberalization. From a practitioner's perspective, a solution to implausible terms-of-trade effects is to substantially raise the values of σ and thereby weaken the terms-of-trade power. However, raising σ can generate implausibly large trade responses to tariff cuts. While the choice of higher elasticities of substitution is perceived to be important to get a "correct" sign on the welfare predictions, the extremely large predicted responses among source countries can be difficult to justify. Welfare, after all, is unobservable, while trade responses can be observed *ex post*. The consumers of trade policy research are frequently policy makers who will focus on predicted trade responses as a means of model assessment.

Single-country trade models parameterize reduced form export supply to the rest of world (and import supply from the rest of the world) rather than modeling a worldwide system of production and consumption. The optimal tariff argument suggests that the choice of these elasticities is critical for determining the welfare effects of tariff changes. While many single-country models maintain a small country assumption (i.e. import supply from the rest of the world is perfectly elastic) the econometric evidence rejects this notion. Broda *et al.* (2008) estimate export supply elasticities and provide evidence consistent with that view that even small countries are able to affect import prices. Further, prior to joining the World Trade Organization (WTO) these countries set higher tariffs rates on inelastically supplied goods, consistent with the Bagwell and Staiger (1999) theory of the General Agreement on Tariffs and Trade (GATT).

As it is conceptually and analytically straightforward, the formulation of import demand found in Equation (18.2) is used in many contexts. CGE modelers might use estimates of σ to measure the response of aggregate imports to the aggregate import price index, the purchase of imports relative to domestically produced goods given a change in relative prices within an industry or substitution between multiple foreign sources of the same narrowly defined goods. This requires only a redefinition of the prices and quantities in Equation (18.2). However, the response of traded quantities to price changes may look quite different across those applications.

Ruhl (2008) discusses differences in the views of σ across two sets of applications. In the macroeconomic real business cycle literature, values of σ in the range of 1−2 are typically chosen to represent aggregate import demand. In the trade policy modeling literature much larger estimates of the elasticity of substitution, in the range of 4−15, are often chosen to represent more disaggregated import demands. Why does the same parameter have different consensus values in the two fields, even in the context of a common Armington representation of behavior? One possibility is that the models differ in the frequency and the persistence of the modeled shocks.

In the international real business cycle literature, low values of elasticity are needed in order to hit macroeconomic calibration targets such as the observed volatility in terms of trade and a negative relationship between terms of trade and the trade balance. The relative price shocks in this framework are often frequent and transitory; exchange rate

shocks are an important source of external shocks. In this context, low values of σ are appropriate because they suggest that demand-side responses to these transitory shocks are likely to be limited.

CGE applications more typically consider long-run responses to a permanent policy shock. In this context, the elasticity of substitution aims to capture the long-run demand response to the policy change. Substitution possibilities are larger over the longer term, so the elasticity of substitution needs to be larger in order to fully summarize demand-side responses.

This suggests that econometric estimates of σ are most useful as model inputs if the econometric data experiment matches as closely as possible the thought experiment conducted by the CGE model. Providing a better understanding of these econometric data experiments is a principal goal of the next two sections.

18.3 IMPORT DEMAND ELASTICITIES

In this section we describe the econometric literature that estimates the price elasticity of import demand. There are many studies that estimate import demand functions in one form or another, including an enormous literature estimating "gravity models" of trade. Relatively few of these studies directly estimate price elasticities, instead focusing on non-price correlates of trade or proxy variables meant to capture trade costs such as distance or whether partners speak a common language. Accordingly, we focus on a set of papers that generate price elasticity parameters for well-known CGE models, listed in Table 18.1, and additional papers that provide useful perspectives on identification issues.

Parameter estimates vary considerably across the literature. Our goal in this section is to provide insights for practitioners as to where these estimates come from so that they can judge which are appropriate for their particular settings. We emphasize how differences in parameter estimates depend crucially on three factors: what parameters are being identified, the nature of the price variation used to identify the parameters and the possibility that the parameters are not being properly identified in the econometric work.

We begin with papers that estimate price elasticities using data on substitution between imports and domestically produced goods within the same industry. These papers typically use time-series data for a single importer and until recently were the primary source of elasticity estimates used in the CGE literature. We argue that these papers suffer from significant data measurement and identification problems and consider time horizons of import response which are ill-suited for use with policy experiments.

We next consider papers that estimate price elasticities using data on import substitution between multiple foreign sources of goods. These papers use cross-sections or panel data, often involving multiple importers. This reduces the role of measurement

Table 18.1 Elasticity estimates in prominent trade policy-focussed models

Model	Structure	Range of elasticities	Econometric study	Econometric technique
Multicountry models				
Global Trade Analysis Project (GTAP)	Armington	$\sigma_D \in [0.9, 17.2]$; $\sigma \in [1.8, 34.4]$	Currently: Hertel *et al.* (2007); previously: Jomini *et al.* (1994); Alaouze *et al.* (1977)	Bilateral cross-section; time series
Michigan Model	Monopolistic competition	$\sigma_D \in [1.02, 5.71]$	Shiells *et al.* (1986)	Time series
Single-country models				
US International Trade Commission (USAGE)	Armington	$\sigma_D \in [1.0, 5.0]$	Gallaway *et al.* (2003)	Time series
MONASH model (Australia)	Armington	$\sigma_D \in [0.9, 17.2]$; $\sigma \in [1.8, 34.4]$	Currently: Hertel *et al.* (2007); previously: Jomini *et al.* (1994); Alaouze *et al.* (1977)	Bilateral cross-section; time series

σ_D represents the demand-side elasticity of substitution between domestic and foreign varieties. σ represents the elasticity of substitution amongst foreign varieties. GTAP follows the "Rule of Two" convention so that $2\sigma_D = \sigma$. The associated econometric study is typically a benchmark study that may or may not supply all the relevant parameters employed in the model.

error and address identification issues by controlling (or instrumenting) for export supply shocks in order to isolate demand parameters. These estimates are in increasing use, both in prominent CGE models and in a larger theoretical and econometric literature in trade that requires data on elasticities of substitution.

Finally, we draw out lessons for the CGE practitioner about which estimates to use in which instances, and highlight some additional concerns related to time horizons, aggregation, and external validity.

18.3.1 Time-series estimates

18.3.1.1 Home–foreign substitution: time-series estimates

Modern multicountry, multiproduct CGE trade models typically contain a triple-nested utility function:

$$\begin{aligned} U &= (Q_1, Q_2, \ldots Q_K) \\ Q_k &= \left(b_{k\mathrm{H}} Q_{k\mathrm{H}}^{\theta_k^{\mathrm{D}}} + b_{k\mathrm{F}} Q_{k\mathrm{F}}^{\theta_k^{\mathrm{D}}}\right)^{1/\theta_k^{\mathrm{D}}}, \quad \theta_k^{\mathrm{D}} = \left(\sigma_k^{\mathrm{D}} - 1\right)/\sigma_k^{\mathrm{D}} \\ Q_{k,\mathrm{F}} &= \left(\sum_c b_{kc} q_{kc}^{\theta_k}\right)^{1/\theta_k}, \quad \theta_k = (\sigma_k - 1)/\sigma_k. \end{aligned} \tag{18.4}$$

The top tier aggregates over $k = 1, \ldots, K$ distinct goods or sectors (food, electronics, transportation, and so on). Depending on the application this can be Cobb–Douglas, CES or a non-homothetic form, for instance where it is desirable to examine income effects. The middle tier aggregates over quantities of home (H) versus foreign (F) sources of goods within sector k and is most commonly written as a CES aggregator. Finally, the bottom tier aggregates quantities over multiple sources c of foreign goods within sector k and is treated as CES or its limiting case of perfect substitutes.

The lower two nests contain preference weights specific to each source. These are most simply thought of as deep parameters in the utility function and play a critical role in calibrating observed trade flows to the data. However, in more complex market structures, these preference weights are endogenous and can be replaced or augmented by the number and/or quality of varieties produced in the source country. We discuss this in Section 18.5 on structural estimation.

The early econometric efforts to estimate import demand elasticities focus on the middle tier and estimate substitution between home and foreign sources of supply. This implicitly treats all sources of foreign goods as perfect substitutes, taking the limit as $\sigma_k \to \infty$ in the bottom nest. Writing out the demand for goods from source $s = \mathrm{H,F}$ in sector k we have:

$$Q_{ks} = (b_{ks})^{\sigma_k^{\mathrm{D}}} \left(\frac{p_{ks}}{P_k}\right)^{-\sigma_k^{\mathrm{D}}} E_k, \tag{18.5}$$

where:

$$P_k = \left(b_{k\mathrm{F}}^{\sigma_k^{\mathrm{D}}} p_{k\mathrm{F}}^{1-\sigma_k^{\mathrm{D}}} + b_{k\mathrm{H}}^{\sigma_k^{\mathrm{D}}} p_{k\mathrm{H}}^{1-\sigma_k^{\mathrm{D}}} \right)^{-1/\sigma_k^{\mathrm{D}}}.$$

P_k is the CES price index over the home and foreign sources, $p_{k\mathrm{H}}$ measures the price of goods in the domestic market and foreign prices $p_{k\mathrm{F}}$ include all trade costs.

The parameter of interest is σ_k^{D}, the elasticity of substitution between home and foreign sources of sector k goods. Alaouze *et al.* (1977), Reinert and Roland-Holst (1992), and Gallaway *et al.* (2003) write about the demand for imports relative to the demand for domestic production and take logs to get:

$$\ln \frac{Q_{k\mathrm{F}}}{Q_{k\mathrm{H}}} = \sigma_k^{\mathrm{D}} \ln \left(\frac{b_{k\mathrm{F}}}{b_{k\mathrm{H}}} \right) - \sigma_k^{\mathrm{D}} \ln \left(\frac{p_{k\mathrm{F}}}{p_{k\mathrm{H}}} \right), \tag{18.6}$$

Note that the homothetic nature of the CES function implies that expenditures and the price index for sector k drop out of this expression. This makes it possible to estimate the slope of the home versus foreign relative demand curve within sector k without worrying about the functional form in the upper level nest. That is to say, expenditures on sector k may or may not be endogenous to prices, but they will not affect relative demands for home versus foreign goods. Similarly, estimating relative demands excuses the econometrician from constructing an appropriate price index for sector k, which requires knowledge of the precise parameter σ_k^{D} we want to estimate.

To estimate (18.6), many papers take the preference weights as exogenous constants, and assume shocks to relative prices are exogenous to changes in quantity demanded. Time-series variation in relative quantities and relative prices is used to identify σ_k^{D}:

$$\ln \left(\frac{Q_{k\mathrm{F}t}}{Q_{k\mathrm{H}t}} \right) = \alpha_k - \sigma_k^{\mathrm{D}} \ln \left(\frac{p_{k\mathrm{F}t}}{p_{k\mathrm{H}t}} \right) + u_{kt}, \quad \text{where } \alpha_k = \sigma_k^{\mathrm{D}} \ln \left(\frac{b_{k\mathrm{F}}}{b_{k\mathrm{H}}} \right). \tag{18.7}$$

A typical estimation would focus on a single importer and estimate separate relative demand curves for as many products as possible given data constraints. Alaouze *et al.* (1977) use quarterly Australian import data for 1968–1975 with 46 commodities aggregated at the four- digit Australian SIC level.[7] Shiells *et al.* (1986) use annual US imports data from 1962–1978 for 41 three-digit SIC industries (122 after aggressive imputation to fill out missing price values). Reinert and Roland-Holst (1992) use quarterly US import data from 1980–1988 aggregated into 163 four-digit SIC sectors, but construct their import price series using information on prices at the TSUSA (Tariff Schedule for US, Annotated) seven-digit level. Gallaway *et al.* (2003) use monthly US import data from 1989 to 1995 aggregated into 309 four-digit SIC sectors, but construct their import price series using information on prices at the HS (Harmonized System)

[7] Alaouze *et al.* (1977) construct Fisher price indices at the tariff-line level, but do not provide details.

10-digit level. The papers that employ quarterly or monthly series also typically include dummies to absorb persistent seasonality.

One challenge in this literature is the very small number of observations available to these econometricians. Since each commodity represents a separate regression, parameters are identified from (at most) 16 annual data points for Shiells *et al.*; 24 and 36 quarterly data points for Alaouze *et al.* and Reinert and Roland-Holst, respectively; and 94 monthly data points for Gallaway *et al.* in total. The numbers at the higher end of the range arrive at this only by using high frequency variation, which may allow for fundamentally different sorts of adjustments in economic variables than longer run changes in prices.

Equation (18.7) is the most common form of the estimating equation, but other variants appear in the literature. Shiells *et al.* (1986) estimate an equation similar to (18.5) using only import quantities as a function of home and foreign prices (introduced separately rather than as a ratio) and expenditures on sector k. Their motivation is to consider a demand function more general than CES, to separately estimate own and cross-price effects, and to examine the total derivative of imports with respect to changes in prices including any expenditure effects. This is a potentially useful approach for CGE practitioners who want to consider more general functional forms in their modeling. Incorporating income effects may be of particular interest when considering trade price shocks that also correspond to sharp changes in the business cycle, such as occurred during the Great Trade Collapse of 2008.

Alaouze *et al.*, Shiells *et al.* and Gallaway *et al.* all incorporate lagged values of the dependent variable, motivated by a model of stock adjustment. This allows imports to respond flexibly to a price shock, both contemporaneously and in subsequent periods. The authors generally find that long run (i.e. lagged) responses to the shocks are somewhat larger than the short run (i.e. contemporaneous) responses. Note, however, that the econometric implementation of short versus long run varies substantially across studies depending on whether observations are monthly, quarterly or annually. A "contemporaneous" trade response at annual frequencies may occur after a three period lagged response in monthly data. In contrast, studies employing cross-sectional and panel variation implicitly contemplate much longer adjustment periods.

18.3.1.2 Problems with time-series estimates: measurement error and simultaneity

Time-series estimates of home–foreign substitution typically find low price elasticities of import demand. Where statistically significant, point estimates are generally around −1 and many estimates are much smaller than this. We next discuss some likely reasons why these papers find very low elasticities and why it is plausible to think that estimates significantly understate the own-price elasticity of demand. The key issues involve measurement error in prices, the construction of the dependent variable and classic problems of simultaneity.

18.3.1.2.1 Attenuation bias: measurement error in prices

If prices are measured with error, estimates of the price elasticity of demand will be attenuated, or biased toward zero. To understand this problem better, consider how the data for estimating Equation (18.7) are typically constructed. Domestic prices are generally constructed using producer price indices at industry level reported by national statistical agencies. One would like a similarly constructed data series on import prices, rigorously built by price sampling common items over time, but these are not available at the level of disaggregation or with the length of coverage necessary for estimation.

Authors instead employ unit values constructed from data on import value and quantity. The numerator is the total value of imports in sector k, arrived at by summing the value of imports in sector k across each source country i, and each disaggregated product category c within the larger aggregate k. Shiells *et al.* (1986) calculates aggregate quantities in a sector with a similar summation, then measures the price as total value divided by total quantity:

$$p_{k\mathrm{F}t} = \frac{\sum_{c\in k}\sum_i p_{ict}q_{ict}}{\sum_{c\in k}\sum_i q_{ict}} = \sum_{c\in k}\sum_i p_{ict}s_{ict}. \tag{18.8}$$

The sector price at time t is a share-weighted average of all the prices within the broader category, where s_{ict} is the quantity share of product c and source i in purchases for sector k.[8] A problem with aggregated unit values is that the quantity measures reported in the data are specific to individual product categories c and can differ across products within an industry summation. A sector like transportation equipment may then aggregate dissimilar units (numbers of cars plus numbers of trucks plus kilograms of tires).

Reinert and Roland-Holst (1992) and Gallaway *et al.* (2003) resolve the mixed unit problem by constructing unit values for each product category c by first aggregating values and quantities over all sources i, $p_{ct} = \sum_i p_{ict}q_{ict}/\sum_i q_{ict}$. These category-specific unit values are then aggregated across product categories c within an industry k using a fixed weight corresponding to the value share of category c in the base year:

$$p_{k\mathrm{F}t} = \sum_{c\in k} w_{c0}(p_{ct}/p_{c0}).$$

Significant measurement error may remain because the quantity measures are themselves notoriously noisy.[9] A consequence is that time-series volatility in unit values far exceeds the corresponding volatility in properly constructed import price indices.

[8] This creates a conceptual mismatch between theory and data because the construction of the price and quantity measures is inconsistent with the assumptions used to aggregate over sources and product categories. These authors treat all source countries as perfect substitutes in order to collapse the lowest tier in equation into a simple summation. However, a consumer that regards a set of goods as perfect substitutes does not look at the average price in that set, but rather the lowest price.

[9] In a recent literature on quality and trade that seeks to explain unit values, the authors put substantial effort into resolving this problem. This literature finds omnipresent outliers, reporting, for example, a preponderance of quantity = 1 observations (see, e.g. Schott, 2004).

This problem is exacerbated when employing monthly or quarterly frequency data at extreme levels of product disaggregation because exporting countries may report sales in only a few periods each year. This means that the product category c average price can move substantially over time due to compositional change in the set of exporters comprising the aggregate.

18.3.1.2.2 Non-classical measurement error in the dependent variable

The potential for measurement error in prices also creates the likelihood of non-classical measurement error in the dependent variable. Suppose we accurately measure the value of imports, M, and noisily measure the quantity of imports $\hat{Q} = Q \cdot e$, where e is the error in measuring true quantities Q. From this we construct prices (unit values)[10] as $\hat{p} = M/\hat{Q} = p/e$. The estimating equation is now:

$$\ln Q_t + \ln e_t = \beta(\ln p_t - \ln e_t). \tag{18.9}$$

As prices are constructed as a function of noisily measured quantities, the error term also appears on the right-hand side, entering with a negative sign. Now, suppose that time-series variation in $\hat{Q}$, $\hat{p}$ comes only from the error term, i.e. $\Delta \ln Q = \Delta \ln p = 0$. In the time series this equation becomes $\ln e_t = \beta(-\ln e_t)$ and estimation would yield an elasticity of -1. That is exactly the value typically found in most time-series estimates![11]

A distinct source of measurement error enters through construction of the denominator of (18.7), domestic consumption of domestic production. Reinert and Roland-Holst (1992) and Gallaway *et al.* (2003) use the value of domestic output less exports. By expressing imports relative to this residual measure we econometrically impose a constraint that price shocks affect US imports and exports with a similar sign. Consider, for example, price changes induced by an appreciation of the dollar. The appreciation should lower the relative price of imports and raise import quantities. However, we would also expect the appreciation to make US goods more expensive on foreign markets and lower exports. Holding output fixed this will cause the denominator to rise and offset the rise in imports. In estimates, this will show up as a muted or potentially even wrong-signed relative quantity response to relative price changes.

Of course, over some horizon we would expect that the quantity of domestic output should adjust so that in the full general equilibrium, we should see the appropriate response. Our point is that these shocks are not simply tracing out movements along a relative demand curve. Rather, the responses depend on the relative elasticity of imports and exports to exchange rate changes and the rate at which output versus trade quantities adjust. Whether this has anything to do with the utility parameters in Equation (18.4) is unclear.

[10] An equivalent problem is created when authors construct prices from error ridden measures of quantity at the disaggregated level, then calculate the quantity index as import value divided by an error ridden price index.

[11] Of course, domestic and import quantities and import prices are also moving around over time so the estimation is not quite this tautological, but unless Q and p move a lot relative to the error term, the estimate will be biased toward -1.

18.3.1.2.3 Simultaneity

Time-series estimates of (18.7) purport to identify the price elasticity of import demand by assuming away the classic problem of supply—demand simultaneity. That is, they treat shocks to supply prices as uncorrelated with the error terms in the demand equation. Consider a few reasons, in this trade context, to be skeptical about these identifying assumptions.

Suppose the home versus foreign relative supply curve is upward-sloping, i.e. relative prices on the right-hand side of (18.7) are a function of the relative quantities consumed. This will be the case in any model in which production possibilities, either across goods or across destination countries, are concave. A shock to relative demand for imports affects supply prices and induces a correlation between prices and the errors in (18.7). This concern is often associated with importers that are "large" on world markets and so affect foreign prices. However, because the demand equation is written in relative terms the same point goes through even for small countries. Even if the country is not large enough to affect world prices, it is presumably large enough to affect its own domestic prices.

In more fully articulated models of the supply side the "taste" parameters $b_{k\mathrm{H}t}/b_{k\mathrm{F}t}$ terms reflect choices made by supplying firms. For example, these taste parameters may reflect the underlying quality of home versus foreign goods. If quality is costly to supply this induces positive correlation between prices and the errors in (18.7), and biases estimates of σ_k^{D} toward zero. This possibility is strongly confirmed in recent econometric work that examines prices and market shares using detailed trade data. Export prices are higher for more capital-abundant and human-capital-abundant countries, and countries with higher prices have larger, not smaller, market shares, especially when selling to high income destinations.[12]

As a related point we can treat Equation (18.5) as the demand facing a single firm and interpret national trade data as an aggregation of all firms engaged in trade. In this case the relative demand for home versus foreign goods depends on the number of firms selling to the importer. With symmetric firms we can write the "taste" parameters in Equation (18.7) as:

$$\alpha_{kt} = \ln \frac{N_{k\mathrm{F}t}}{N_{k\mathrm{H}t}} + \sigma_k^{\mathrm{D}} \ln \left(\frac{b_{k\mathrm{F}t}}{b_{k\mathrm{H}t}} \right). \tag{18.10}$$

If firm entry into the importing market is time varying and correlated with prices, as in the heterogeneous firm literature discussed in Section 18.4, we have an additional source of simultaneity bias in estimated elasticities.[13]

[12] See Schott (2004), Hummels and Klenow (2005), Hallak (2006), Choi *et al.* (2009), Khandelwal (2010), and Hallak and Schott (2011).

[13] However, the direction of the bias is not obvious in this case, and depends on whether high prices are negatively or positively correlated with entry. In models of free entry with pure horizontal differentiation, low cost producers will tend to have more varieties (see Melitz, 2003), whereas in models that combine horizontal and quality differentiation, high-quality (high-cost) providers have more varieties.

18.3.2 Foreign–foreign substitution: cross-sectional and panel estimates

A more recent literature estimating import demand elasticities has addressed the simultaneity problem in time-series estimates using two distinct approaches: controlling for unmeasured shocks to supply and instrumenting for prices. However, in order to generate the variation necessary to address these problems this literature focuses on a different parameter – the elasticity of substitution across multiple foreign sources of goods.

Foreign–foreign substitution appears in the lower tier of Equation (18.4). Substitution across multiple foreign sources is especially important for modern multicountry models, especially those focused on policy shocks such as preferential trade agreements that favor one partner relative to another. Early on in the development of multicountry CGE models, estimates of foreign–foreign substitution were not available and so modelers employed *ad hoc* assumptions. As an example, in early versions of GTAP, estimates of home–foreign substitution σ_k^{D} were drawn from the time-series literature reviewed in Section 18.3.1 and based on these values, foreign–foreign substitution was constructed using the "Rule of Two," or $\sigma_k = 2\sigma_k^{\mathrm{D}}$. More recently, direct estimates of σ_k in the lower level nest have been obtained and these show a marked difference relative to purely time-series estimates. We discuss these next.

18.3.2.1 Exploiting price variation induced by trade costs

Our discussion of this problem begins with Hummels (2001) and Hertel *et al.* (2007). Import demand is similar to that in Equation (18.5) except that the subscripts now refer to a particular exporter i and importer j. We further augment the supply side to include the possibility of quality differences across suppliers, as well as differences across markets in the number of firms exporting. Suppressing commodity k subscripts for ease of notation, the augmented equation is then:

$$q_{ij} = n_i(b_i)^{\sigma}\left(\frac{p_{ij}}{P_{j\mathrm{F}}}\right)^{-\sigma} E_{i\mathrm{F}}. \tag{18.11}$$

This expression uses the nesting structure of the upper level model so that expenditures on foreign goods and the price index are specific to an importer. In principle one could estimate this expression in precisely the same manner as that used in the time-series analysis of home versus foreign purchases. Alternatively, one could exploit the rich data variation available in bilateral trade data to estimate the equation in the cross-section.

To see how this works, we write the price of exporter i's good sold in market j as $p_{ij} = p_i(1 + \tau_{ij})$, where τ_{ij} includes tariffs and transportation costs. That is, factory gate prices p_i charged by exporters are invariant to the trade costs added to them before the goods reach their destination. To be clear, trade costs might affect factory gate prices in

general equilibrium through mechanisms such as factor price changes, but once p_i is determined in a period, it is the same for all destinations. The result that bilateral variation in trade costs does not pass-through into bilaterally varying factory gate prices falls directly from a combination of monopolistic competition on the supply side, CES preferences (so that trade costs do not change the elasticity of demand facing the firm) and "iceberg" trade costs.[14]

Substituting in for prices and taking logs we have:

$$\ln q_{ij} = \ln n_i b_i^{\sigma} - \sigma \ln p_i + \sigma \ln P_{j\mathbb{F}} + E_{j\mathbb{F}} - \sigma \ln\left(1 + \tau_{ij}\right). \tag{18.12}$$

Quantities can be problematic, for reasons noted above, and so this equation can be rewritten in value terms by multiplying both sides by the price. This change to a better-measured dependent variable changes nothing in the estimating equation except the implied coefficient on price becomes $1 - \sigma$.[15]

Above we noted the problem with employing prices in the demand equation: they are measured with error and they are potentially correlated with other supply characteristics, $\ln n_i b_i^{\sigma}$. This problem can be solved by making use of the extensive variation available in bilateral trade flows. Note that all variables except for trade costs are denoted either with an importer j or an exporter i subscript. If we have two or more exporters for each import destination, and two or more importers for each export source we can eliminate these variables by differencing.

There are several distinct approaches to this in the literature. Hummels (2001) and Hertel *et al.* (2007) use a single cross-section of commodity level bilateral trade – exports from every country worldwide into a subset of importers (the US and six Latin American countries) that report detailed tariff and transportation cost data. Including vectors of exporter α_i and importer α_j fixed effects they estimate[16]:

$$\ln pq_{ij} = \alpha_i + \alpha_j - \sigma \ln\left(1 + \tau_{ij}\right) + e_{ij}. \tag{18.13}$$

Note that the fixed effects have a particular interpretation in terms of the underlying parameters of the demand equation:

$$\alpha_i = \ln n_i b_i^{\sigma} + (1 - \sigma) \ln p_i$$

$$\alpha_j = \sigma \ln P_{j\mathbb{F}} + E_{j\mathbb{F}}.$$

14 The result does not go through when employing an alternative demand structure with variable markups such as quadratic or translog utility or when using trade costs that are additive rather than multiplicative with prices. We address this point when considering external validity below.

15 In principle, time-series estimators could use the same transformation, but do not because they typically estimate a unitary price elasticity. This implies that expenditures are invariant to prices.

16 Suppose that there are 100 importers and 100 exporters of a particular good. This translates to 200 parameters to estimate and 99,900 distinct trade flows (excluding each country's purchases from itself).

These fixed effects could be estimated directly (and the information in them preserved and interpreted structurally), or eliminated by subtracting out importer and exporter means.

Equation (18.13) sidesteps problems with simultaneity and price mis-measurement by assuming that price, quality and variety are the same in each destination market that exporter *i* sells to, and can therefore be absorbed by the fixed effect. In essence, the fixed effect holds constant the position of the demand curve across all possible markets. Other things being equal, exporters who provide a lot of variety at high quality and low prices will have large market shares.

With exporter supply characteristics fixed, Equation (18.13) uses bilaterally varying trade costs instead of prices to identify the price elasticity of demand. That is, variation in trade costs across import markets provides the price variation necessary to trace out the slope of the demand curve. There are differences across import markets in expenditures (due to size or idiosyncrasies in demand) and price indices (due to the intensity of competition), but these differences are absorbed by the importer fixed effects.

A closely related approach is used by Romalis (2007) who estimates the import demand elasticity by attributing relative changes in imports to changes in relative trade costs during the implementation of the North American Free Trade Agreement (NAFTA). He constructs a difference-in-difference estimator based on an equation like (18.11). Consider US imports of a particular product from two distinct sources (Mexico and a reference country such as Korea) in 1990. Expressing these as a ratio we eliminate all variables that are specific to the US as an importer. This can be done for any other importer as well, and so Romalis constructs a similar ratio for EU imports from Mexico and the reference country. Subtracting the EU import ratio from the US import ratio, we eliminate all variables that are specific to Mexico and Korea as suppliers. What this leaves is time-varying changes in NAFTA import tariffs for the US. That is, US tariffs on Mexico are dropping relative to Korea and the same is not happening for EU imports. This should generate a differential growth rate in imports determined by the import demand elasticity.

If we compare the Romalis (2007) approach to Hummels (2001) and Hertel *et al.* (2007) there are three clear differences. The Romalis approach yields a precisely estimated elasticity from a data experiment that is conceptually very close to that used by many modelers. There is a tariff cut that affects some but not all suppliers and the estimates examine how elastic is the import response to this cut holding fixed all other factors in the model. Hummels and Hertel *et al.* exploit data in cross-section, so there is no element of a time-series trade response to changing prices. Instead, the data experiment should be understood as a kind of long-run trade response to long-standing trade costs.

A second point of contrast is that Romalis (2007) has relatively little variation in the tariff changes to exploit and so pools over all products in order to use cross-product

variation in the depth of tariff cuts. This yields a single elasticity for all products rather than a set of estimates that are product specific. Hummels and Hertel *et al.* have much more bilateral variation in trade costs due to both preferential tariffs and transportation costs, and so are able to estimate elasticities for many products.

Finally, while Hummels and Hertel *et al.* pool over all importers to estimate a single elasticity for a given product, Romalis' approach requires that he choose specific comparison and reference countries. As he varies the countries chosen, parameter estimates vary, sometimes substantially. In a demand system where all varieties are symmetric substitutes, one should find symmetric quantity responses to price changes. This raises the possibility that demand systems with variable elasticities or asymmetric cross price effects better represent the data. We return to this point in Section 18.3.5.

The time-series literature estimates elasticities that are typically around −1.0 or less. Hummels estimates elasticities under a variety of pooling and aggregation assumptions, with mean estimated elasticities ranging from −5.3 (for SITC two-digit) to −7.3 (for SITC four-digit). Hertel *et al.*, using trade data concorded to 39 GTAP categories, find median elasticity estimates of −6.55 (−7.0 mean). Romalis pools over all product categories but reports many estimates that vary by country sample and controls employed. The median across these estimates is −6.9.

Compared with the time-series literature, these are enormous differences. To see this, consider the effect of cutting tariffs. Romalis reports that in 2000, the simple average most favored nation (MFN) tariff rate imposed by the US was 5.2%, and nearly 0% for Canada and Mexico as a result of Canada−US Free Trade Agreement and NAFTA agreements. Using estimates from the time-series literature, we would calculate that cutting tariffs from 5.2 to 0% would generate 5.2% more trade. Using estimates from the cross-section/panel literature we would calculate that the same tariff cut would generate from 28 to 38% more trade.

18.3.2.2 Problems with using trade costs in cross-section/panel

There are some limitations to approaches that difference or dummy out nuisance parameters and identify price elasticities off of trade costs. First, it can only be used where there is bilateral variation in *ad valorem* trade costs sufficient to identify the price elasticity. Hummels (2001) and Hertel *et al.* (2007) use cross-sectional variation in transportation costs and preferential tariffs, with much of the identification coming from transportation costs. Romalis (2007) uses panel variation in the implementation of preferential tariffs. Transportation costs are not available for many countries and there are limited examples of preferential tariff cuts of the sort exploited by Romalis. As a consequence these exact estimation techniques can only be used for certain countries or certain time periods. Whether the resulting estimates have external validity, i.e. whether they are useful for other countries and time periods is a question we take up in Section 18.3.5.

Are there other trade costs that might be employed to extend the data samples? MFN tariffs provide an *ad valorem* shifter of export prices, and so could potentially identify price elasticities. The problem is that, by definition, they do not vary across exporters for a given importer. The use of importer fixed effects as in Equation (18.13), or differencing as in Romalis, eliminates this usable variation. Some authors have focused on distance between markets as a trade cost proxy. This variable does have bilateral variation but because it is not in *ad valorem* equivalent form it cannot be used directly to calculate the price elasticity. Conceptually, we could model *ad valorem* trade costs in (18.13) as $\ln(1+\tau_{ij}) = c + \delta \ln DISTANCE_{ij}$. Replacing *ad valorem* costs with distance in Equation (18.13) yields a regression coefficient of $-\sigma\delta$. Extracting σ from the overall effect requires information on δ, the elasticity of trade costs with respect to distance.

A second problem with the trade cost approach is that it assumes that trade costs themselves are uncorrelated with the error terms in the bilateral trade Equation (18.13). A large literature on the political economy of protection suggests that tariffs will be highest when potential import penetration is greatest.[17] Another literature shows that *ad valorem* transportation costs depend on the scale of trade, as high trade volumes lead to procompetitive entry by shippers and as trading countries respond to high volumes with infrastructure improvements that lower costs. Further, since transport costs are proportional to quantities rather than values shipped, high-priced items enjoy lower *ad valorem* costs.[18] In these cases, the concern is that trade costs are endogenous to quantities traded.

This is a potentially serious drawback and so some care must be taken to specify the source of the endogeneity and whether trade costs are correlated with the error terms after conditioning on other model variables. For example, the political economy literature highlights the possibility that, when looking across industries, levels of protection are correlated with levels of import penetration. It is unclear whether there should be a similar concern with the variation that Romalis exploits — variation in preferential tariffs across trading partners within a given industry.

Similarly, transportation costs depend on product prices and the scale of trade (via entry and infrastructure quality), which would seem problematic for the approaches used by Hummels and Hertel *et al.* However, these effects are likely to be captured by the inclusion of importer and exporter fixed effects. Put another way, the fixed effects estimators are designed to eliminate nuisance variables $n_i b_i, \ln p_i, P_{j\mathrm{F}}$ and $E_{j\mathrm{F}}$ in (18.12) in order to cleanly identify the slope of the import demand curve. Along the way they also succeed in eliminating the obvious source of endogeneity between trade costs and quantities traded, leaving trade costs uncorrelated with the errors in Equation (18.13).

17 Some examples include Trefler (1993), Grossman and Helpman (1994), and Goldberg and Maggi (1999).
18 See Hummels and Skiba (2004) and Hummels *et al.* (2009).

18.3.3 Instrumental variables approaches to identifying foreign–foreign substitution

The next set of papers we discuss identify import demand elasticities using export prices for multiple exporters. This is similar to the older time-series literature discussed in Section 18.3.1, except that these papers address mis-measurement of prices and simultaneity using an instrumental variables approach. Like the literature that uses trade costs in Section 18.3.2, they estimate foreign–foreign rather than home–foreign substitution.

A general problem in the estimation of structural parameters is the difficulty that arises when both supply and demand parameters must be separately identified from equilibrium data. Equilibrium prices and quantities lie on both the supply and demand functions, and it can be difficult to identify the slopes of either function when both supply and demand shocks are present. This is a longstanding problem within the discipline; it is not limited to the international trade literature. A standard solution is to employ an instrumental variables estimator in which particular exogenous shocks are assumed to shift either supply or demand but not both, allowing separate identification of the supply and demand functions.

The difficulty lies in the shortage of appropriate instruments. To properly instrument for prices in the import demand Equation (18.7), we need a variable that is correlated with prices and uncorrelated with the error term in (18.7). An early example of this approach is found in Shiells *et al.* who instrument for world and domestic prices using foreign and domestic factor prices. In that paper, instrumental variables (IV) and ordinary least squares (OLS) regressions yield similar elasticities. This is likely because the aggregated nature of the data results in weak instruments.[19]

Erkel-Rousse and Mirza (2002) use a panel of commodity level bilateral trade flows for the Organization for Economic Cooperation and Development (OECD) and an estimation approach that is similar to Hummels, Hertel *et al.* and Romalis. Starting from Equation (18.12) they difference across exporters to remove all variables that are specific to each importer-year, including expenditures, price indices and MFN tariffs. This leaves prices which vary both over exporters and across time to identify the demand elasticity. Erkel-Rousse and Mirza (2002) instrument for prices using wages and exchange rates in order to address identification concerns related to measurement error and simultaneity bias. This reveals something very interesting relative to the early papers. When they use OLS they find the elasticity of imports with respect to price changes is −0.83, very similar to the older literature. However, when they instrument for prices and remove measurement error and simultaneity bias, their estimate jumps to −3.75 (for IV) and

[19] With weak instruments, IV estimates are biased toward OLS. The authors are using time series with as little as six observations to explain unit values that are aggregated over both countries and products using factor prices from individual countries. While the paper does not report first stage diagnostics, it would be surprising if this resulted in a strong instrument.

−7.58 [using generalized method of moments (GMM)]. These numbers are quite similar to the literature that uses trade costs in cross-section or panel to identify the elasticity.

Kee, *et al*. (2008) use a GDP revenue function approach to motivate estimation of import demand elasticities using a translog functional form applied to highly disaggregated data. The translog function is extremely general, but with some simplifying assumptions the actual estimating equations are closely related to those discussed in relation to this point. The expenditure share for good n depends on its log price relative to the average price of all other goods, as well as endowments. Kee *et al*. instrument for prices using trade cost proxies (distance) and average prices for other exporters in the same product category. The latter is likely to be correlated with prices for a specific exporter if production cost shocks for a given product are common to many exporters worldwide. It is less clear why these rest of world prices pass the exclusion restriction, i.e. why they do not belong directly in the demand equation. This is, perhaps, an appeal to the underlying demand system. In a CES demand function competing prices for similar goods would clearly belong in the demand equation via the price index, while in the translog, these are just a small number of prices relative to the many thousands that affect overall demand.

Kee *et al*. use UN Comtrade data on imports for 4900 HS six-digit products and 117 countries for the years 1988−2001. They estimate over 377 000 elasticities, and report a mean import demand elasticity of −3.12. However, the mean value results from a long left tail; a kernel density shows that most of the estimates lie between 0 and −2. Like the time-series literature these estimates are centered on −1.0. Contrasting this result with the findings in Erkel-Rousse and Mirza, we suspect that relatively poor instruments may be responsible for the lower estimated elasticities.

18.3.4 Systems of equations without instruments

18.3.4.1 The Feenstra method

Among econometrically oriented trade economists, perhaps the best known approach to identifying import demand elasticities is due to Feenstra (1994) and its extension in Broda and Weinstein (2006). This technique is challenging to understand and connect to the rest of the literature surveyed here. We attempt such an explanation here so that practitioners might better understand how, and under what conditions, the estimator works and what this implies for adoption of elasticities arising from these papers.

The starting point is Leamer (1981) who investigates the problem of identifying supply and demand parameters in a simultaneous equations system without instruments. We will translate his notation to a problem similar to the rest of this paper. Suppose we have a supply and demand system as follows:

$$\begin{aligned} \ln q_t &= \alpha + \sigma \ln p_t + \varepsilon_t \\ \ln q_t &= \gamma + \omega \ln p_t + \mu_t. \end{aligned} \tag{18.14}$$

The first equation describes demand, the second supply, so $\omega > 0$, $\sigma < 0$. By assumption the errors in the two equations are uncorrelated. We lack instruments to separately identify supply and demand parameters and so the model is under-identified. Leamer shows that we can use information about the variance in the shocks to price and quantity to describe a relationship between supply and demand elasticities. This relationship is a hyperbola defined by:

$$(\hat{\sigma} - b)(\hat{\omega} - b) = \frac{s_q^2}{s_p^2}\left(r_{pq}^2 - 1\right), \tag{18.15}$$

where b is coefficient in an OLS regression of quantity on price, s_q^2 is the sample variance in quantities, s_p^2 is the sample variance in prices, and r_{pq}^2 is the squared sample correlation with price and quantity.

What is the intuition? In time-series data we see a cloud of price, quantity points that arise from a combination of supply and demand shocks. The shape of this cloud tells us something about the nature of shocks that hit over time. Suppose we see a data cloud of points like the oval "S" in Figure 18.2. There is upward slope in the data cloud, meaning that the sample correlation in p and q is positive. This is consistent with the view that both demand and supply shocks are occurring but that the variance in the demand shocks is greater than the variance in the supply shocks.[20] In Figure 18.3, we have drawn the hyperbola from Equation (18.15) corresponding to data cloud "S". Leamer's insight is that if we assume that supply curves slope up and demand curves slope down we can further bound the possible values of these parameters. Define b_r as the point where the hyperbola cuts the axis, i.e. where $\sigma = 0$ and there is no slope to the demand curve. The

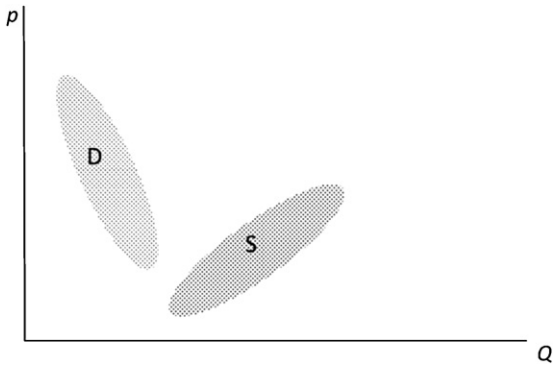

Figure 18.2 Illustrative shock patterns.

[20] Contrast this with the classic IV case in which we use an instrument like income to identify only demand shocks, then trace out a supply curve to yield a single slope. Here we have both kinds of shocks but the demand shocks are dominating.

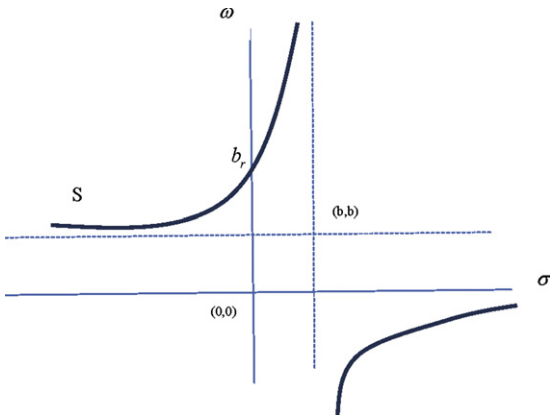

Figure 18.3 A Leamer hyperbola that bounds elasticity estimates.

part of the hyperbola with feasible parameters is given by $\sigma < 0$ and $0 < b < \omega < b_r$. This corresponds to a relatively tight bound on the supply elasticity, but provides virtually no information about the demand elasticity. In other words, an upper-sloping data cloud is consistent with a relatively small set of supply parameters but could be generated by almost any demand slope.

In contrast, suppose we have data cloud "D" in Figure 18.2, so that the variance in the supply shocks is greater than the variance in the demand shocks, and the correlation between price and quantity is negative. This corresponds to the hyperbola D in Figure 18.4, where "b" is negative. Again we can focus on the part of the hyperbola with sensible values. Defining b_r at the point where the hyperbola cuts through $\omega = 0$, we can bound parameters on the hyperbola in the region $b < \sigma < b_r < 0$ and $\omega > 0$. This corresponds to a relatively tight bound on the demand elasticity, but provides virtually no information about the

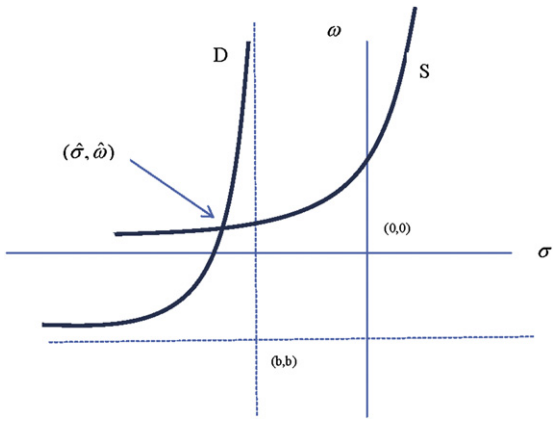

Figure 18.4 The intersection of two Leamer hyperbolae isolates two point estimates.

supply elasticity. A downward-sloping data cloud is consistent with a relatively small set of demand parameters but could be generated by almost any supply slope.

Feenstra (1994) uses this insight and points out that an interesting feature of bilateral trade data is that we potentially have up to N different data series, one for each of the N suppliers to a market. If these suppliers face a different set of demand and supply shocks, we have different hyperbolas to describe the bounds on possible parameters. Suppose we have two suppliers. In the first, the time series reveals a lot of demand variance (cloud D), and in the second, the time series reveals a lot of supply variance (cloud S). This results in two different hyperbolas, one for each country. We show both of these in Figure 18.4. Along each hyperbola are all the possible combinations of elasticities consistent with the corresponding data series. This does not help us if the demand and supply elasticities are different for the two countries. However, if we assume that the supply and demand elasticities are the same for both countries this can happen at only one point in parameter space. That point is given by the intersection of the Leamer hyperbolas, $(\hat{\sigma}, \hat{\omega})$.

Finding this intersection point is the central idea of the Feenstra (1994) estimator. Next, we describe particular implementation details and contrast his approach with the extension in Broda and Weinstein (2006). The starting point is an import demand equation similar to (18.12) but with a few differences: all variables refer to US imports rather than a bilateral cross-section with many importers; all variables have time subscripts; there is a careful adjustment of prices and price indices to reflect growth in variety (the measurement of which is the main focus of the paper); and prices are described in terms of delivered prices, i.e. they are not decomposed into an origin component and a trade cost. In addition, there is a reduced form export supply curve like that in Equation (18.14).

To eliminate unmeasured terms in the import demand equation, Feenstra converts quantities into value shares and first differences the data to give an import demand equation of:

$$\Delta \ln s_{it} = \phi_t - (\sigma - 1)\Delta \ln P_{it} + \varepsilon_{it}.$$

The term ϕ_t captures changes in the price index over time for a given sector arising from the combined changes in variety, quality and adjusted prices for all exporters. To eliminate the price index, Feenstra expresses country i exports relative to a reference country k. This yields a system of supply and demand written in double differences:

$$\tilde{\varepsilon}_{it} = (\Delta \ln s_{it} - \Delta \ln s_{kt}) + (\sigma - 1)(\Delta \ln p_{it} - \Delta \ln p_{kt})$$

$$\tilde{\delta}_{it} = (1 - \rho)(\Delta \ln p_{it} - \Delta \ln p_{kt}) - \left(\frac{\rho}{\sigma - 1}\right)(\Delta \ln s_{it} - \Delta \ln s_{kt}),$$

where ρ depends on σ and ω.

It is useful to contrast this approach with the Hummels (2001), Hertel *et al.* (2007) and Romalis (2007) approaches that use trade cost data to identify demand parameters.

All these papers eliminate importer-specific terms from the estimating equation by differencing exports from country *i* relative to a reference country or a set of reference countries represented with an importer fixed effect.

To eliminate exporter specific shocks arising from unmeasured variety and quality, Romalis differences the US import demand equation relative to another importer such as the EU, while Hummels and Hertel *et al.* use an exporter fixed effect estimated from multiple exporters. These papers take the position that an exporter has the same price, variety, and quality for a particular good in every import market they sell to. In contrast, Feenstra assumes that differences across exporters *i* in quality or variety can be either eliminated via first differencing (e.g. if they are constant over time) or are captured in the error term, which is assumed to be uncorrelated with the error term in the supply equation. Both strands of the literature take a strong stand on the nature of the unobserved terms in the import demand equation — one asserting that differencing across exporters eliminates correlations with supply shocks, the other asserting that differencing over time accomplishes the same goal.

As a final step, Feenstra assumes that the errors in the supply and demand equations are uncorrelated, and multiplies them together to get a single equation in prices and market shares. This is where the connection with Leamer (1981) comes in. For each exporter *i*, we can describe the time series in prices and market shares in terms of the sample variance and covariance in prices and market shares. That is, Feenstra does not preserve the exporter *i*/time *t* variation in the data, but collapses it into variance and covariance terms appearing in Equation (18.15) — one set for each of the *N* countries that export a given product to the US in the sample.

Using this, one could simply construct a Leamer hyperbola for each exporter. If the variance, covariance patterns differ, any two exporters are sufficient to identify the elasticities of supply and demand. The difficulty is that there are $N-1$ such comparisons, and no particular reason to think that all *N* hyperbolas intersect at a single point. Instead of doing a pairwise comparison, Feenstra uses a GMM estimator to extract the average tendency, over the *N* countries, for the variance in prices to be correlated with the variance in quantities and the covariance between prices and quantities. If the regression coefficients fall within certain prescribed ranges, one can then extract the implied values for σ and ω.

Feenstra (1994) illustrates his estimator using a small number of products in US imports. Broda and Weinstein (2006) extend the Feenstra estimator in a wide-ranging study of US imports, estimating tens of thousands of elasticities at varying levels of aggregation and over two sample periods (1972–1988 and 1989–2001). They report median elasticities of 3.7 on their most disaggregated samples, with considerably higher means because of skewness in the distribution of estimates.[21] The median of these

[21] Their estimates are smaller in more aggregated samples and larger for reference priced commodities.

estimates lies between the very low elasticities found in the time-series literature and the considerably higher estimates found in the literature using trade costs to identify price variation. These estimates are provided on the authors website and provide the most comprehensive set of elasticities available.

18.3.4.2 Problems with the Feenstra method

The identification strategy used in Feenstra and Broda—Weinstein is elegant, but it rests on several strong assumptions. (i) Both import demand and export supply elasticities are common across all exporting countries. This assumption is central: Leamer's insight is that without instruments a regression of quantity on price identifies a mapping between supply and demand elasticities. We can only pin down where the parameters lay on two different hyperbolas if they intersect at a (common) set of elasticities.

(ii) Feenstra collapses the i,t variation in the data down to a single exporter specific description of data given by the sample variance, covariance in price and quantity. This assumes that the error process for both supply and demand is stationary and that the number of periods T is sufficient to reach the appropriate asymptotic properties (i.e. that the sample variance is equal to the true variance). This is highly problematic for small T. In many instances, exporters ship a particular product infrequently, which means that the available T is significantly less than the total time span of the data.

(iii) Feenstra assumes that once the data has been first differenced and differenced relative to a reference country, there is no remaining correlation between the errors in the supply and demand equations, and that there is enough year on year variation to identify the relevant parameters. The demand curve is identified if there are exporter i time t specific shocks to supply prices that are uncorrelated with shocks to preferences. There are many possible shocks that fit this bill, corresponding to the instruments used in the rest of the literature. Examples include shocks to trade costs, factor prices, exchange rates, or productivity. The supply curve is identified if there are i,t shocks to preferences that are uncorrelated with supply prices. This is more difficult. The trick is that these taste shocks have to be exporter specific, which rules out many of the usual demand shifters. Importer income, for example, shocks demand for all source countries in the same way and is eliminated by differencing relative to the reference country. Other taste shocks such as changes to variety or quality only work if they are uncorrelated with supply shocks, i.e. new varieties or better qualities are not costly to produce. It is unclear to us what plausible source of taste shocks remains to identify the export supply curve.

Feenstra's technique generates both export supply and import demand elasticities, though the supply elasticity is not a focus of either Feenstra (1994) or Broda and Weinstein (2006). Soderbery (2010) revisits Feenstra's IV estimation technique, using Monte Carlo analysis. He simulates data from a structural parameterization that produces data with strong similarities to Feenstra's data, and evaluates the small sample properties of Feenstra's estimator. Soderbery shows that in samples consistent with what data are

available, the estimator produces structural estimates of import demand elasticities that quickly converge to the appropriate levels, while estimates of the import supply elasticity are biased upward. This is consistent with the view that, even pooling over multiple exporters, there is insufficient year on year variation in exporter-product-specific demand shocks necessary to identify the slope of supply curves.

A final problem relates to implementation of the Leamer hyperbolas. As Figures 18.3 and 18.4 make clear, the hyperbolas pass through implausible ranges of parameter space. Leamer argues that one can eliminate the implausible ranges by simply imposing structure: demand slopes down, supply slopes up. However, there is nothing to guarantee that the intersection of hyperbolas from multiple exporters occur in plausible parameter ranges. Further, because demand and supply parameters are extracted from a square root of the coefficients produced by the GMM procedure, negative coefficient estimates imply structural parameters with imaginary values.[22] This problem motivated Broda and Weinstein to use a grid search method that maximizes a likelihood function only over plausible ranges of parameter space. It is unclear to us whether these grid searched values bear a strong relationship to the points in parameter space identified by the Leamer hyperbolas, and if not, how to interpret the entire enterprise. Given the prominence of these studies, and the great use to which the associated parameters might be put, we think further rigorous study along the lines of Soderbery (2010) is required.

18.3.5 Which import demand elasticities to use? A guide for the practitioner

In the preceding sections we have discussed econometric studies that use very different data samples, response horizons and estimating techniques, and arrive at elasticities as much as an order of magnitude different from each other. This raises the critical question: which elasticities are "right?" Or at least, which are right for the particular modeling application at hand? In this section we summarize some practical guidance to selecting elasticities.

18.3.5.1 Magnitudes

Time-series estimates focused on home–foreign substitution and find elasticities of −1.0 or less. Cross-section and panel estimates focused on foreign–foreign substitution find elasticities equal to −5.0 for the median product, and with some products exhibiting much larger elasticities. This is an enormous difference, with significant consequences for a host of positive and normative questions discussed in Section 18.2. Should practitioners use large or small elasticities? The answer to this turns on two questions: why are these estimates so different and how do modelers plan to use them?

[22] In our experience and from informal conversation with others who have experimented with these procedures, the imaginary number problem can occur in as many as half of the estimates.

Section 18.3.1.2 outlines several reasons to think that time-series estimates are biased downward, and in the case of errors in variables, biased toward the most commonly occurring estimate, −1.0. However, there are (at least) two reasons to think that these differences are, possibly, more than artifacts of estimation.

Strictly speaking, the studies estimate different parameters. Time-series studies estimate home−foreign substitution in the middle tier of the utility function, Equation (18.4), while cross-section and panel studies estimate substitution across multiple foreign sources in the lowest tier of (18.4). Perhaps consumers regard two foreign varieties as being close substitutes, and regard foreign and domestic varieties as more distant substitutes. We are unaware of any evidence that is directly informative on this point, but it is not difficult to tell stories consistent with this.[23] When we treat varieties as symmetric CES substitutes we gloss over a richer characterization of product space that may be present.

The estimated elasticity of import demand may depend on a set of responses that are active at different time horizons. Several papers in the time-series literature exploit data at monthly or quarterly frequencies, while the literature using trade costs to identify price shocks exploits a single cross-section or panel data over a decade with a single discrete change in tariffs. The gap in estimated elasticities is largest between these two groups. A third group of papers use annual time-series data and find elasticities between these extremes. Time-series papers that include lagged adjustment report somewhat higher elasticities than when focused on contemporaneous changes.

What do we make of this? The article that comes closest to comparing these explanations is Erkel-Rousse and Mirza (2002). When they use OLS with an annual time series of prices to estimate substitution across multiple foreign sources they find elasticities of −1.0. Using instrumental variables on the same data boosts the elasticity to −3.75 and GMM raises it to −7.58. This clearly shows that estimates based on higher frequency data can generate much higher demand elasticities when prices are properly instrumented, and conversely that estimating off of foreign−foreign substitution can generate much lower demand elasticities when simultaneity is ignored. We cannot claim this directly provides evidence that home−foreign import demand elasticities are, in fact, much larger and would be identified as such if properly estimated. However, it certainly suggests that price mis-measurement and simultaneity can dramatically lower estimated elasticities in the trade context.

We think this clearly indicates that much larger elasticities are appropriate when employed in trade policy experiments. The challenge lies in knowing what to use in a macroeconomic time-series context. Ruhl (2008) points out the contrast between the

[23] For example, local firms may adapt their products to match local tastes while foreign firms provide less adapted "generic" sets of products. Or differences in supply conditions may cause foreign firms to specialize in different kinds of varieties than domestic firms within a broader industrial sector as a function of comparative advantage.

high elasticities used in the trade policy literature and the much smaller elasticities used in the macroeconomic real business cycle literature. Low elasticities are needed to hit macroeconomic calibration targets such as the observed volatility in the terms of trade and a negative relationship between the terms of trade and the trade balance.

Our view is that these low elasticities are best thought of as reduced forms. They are a reduced form of the real economic adjustments happening in the face of international price shocks that are frequent and transitory. When buyers face price shocks they may be slow to identify new sources of supply, especially if products in question are differentiated. A permanent price shock associated with a permanent change (e.g. due to tariff cuts) might make this search worthwhile. Over longer horizons supply itself can begin to adjust, as firms enter and exit markets, or firms customize products for consumers. Modeling time-series adjustments with small elasticities is then a reduced form way to represent slow adjustment process in the theory rather than writing out these slow adjustment processes explicitly.

Low elasticities are also a reduced form econometrically in the sense that the relationship between import prices and quantities does not properly identify the slope of the import demand curve described in Equation (18.6). For many purposes such as matching quantitative responses to price shocks, treating the low elasticity as if it were actually the slope of the demand curve is reasonable. In other contexts such as evaluating welfare changes associated with price shocks, it is a very bad approximation indeed.

Some recent papers have begun to focus on a related issue, the role of uncertainty over the future course of trade policy. Handley and Limao (2010) and Handley (2011) model exporting behavior when firms are heterogeneous, costs of export market entry are sunk and there are random elements to trade policy. In this context firms respond to policy commitments that reduce uncertainty about future trade policy. This analysis suggests that trade agreements that reduce both bound and applied tariff rates generate larger quantity response than do unilateral reductions in applied rates. In applications to preferential commitments by both Portugal and Australia the authors find that reduced trade policy uncertainty increases trade in a manner that is distinct from the effect of reductions in the average applied tariff.

18.3.5.2 External validity

Most if not all trade-focused CGE work takes import demand elasticities from the literature and assumes they hold for all modeled countries and time periods. The problem is that the econometric literature estimating import demand elasticities relies on older data samples from a fairly narrow and potentially unrepresentative set of countries. Early time-series estimates used imports for either the US or Australia dating back to the 1960s. Feenstra (1994) and Broda and Weinstein (2006) use US imports dating back to the 1970s. Hertel *et al.* use US and Latin American imports in a cross-section in 1994.

Romalis (2007) is more up to date, comparing changes in US and EU imports through 2000. The broadest of these studies is Erkel-Rousse and Mirza (2002) who use all of the OECD in the 1970s and 1980s.

Should practitioners be concerned that they are drawing parameters estimated from a single importer's time-series behavior and applying them globally, or using a particular liberalization episode to estimate parameters that will be applied to all subsequent episodes? In this section we address three principal concerns related to external validity of estimates. We begin by addressing parameters as point estimates with associated standard errors, then discuss challenges related to country samples and levels of aggregation. The purpose is to guide the practitioner in considering whether external validity is likely to be a problem in their context.

18.3.5.2.1 Point estimates and standard errors

Suppose that the underlying parameters, σ_k from the utility function in Equation (18.4), are the same for all countries and time periods. If we can properly estimate the parameter for any one country or time period, we then know it and can apply it broadly. The problem is that point estimates from any particular econometric study are just one draw from the underlying parameter distribution, and the precision of the estimate depends on idiosyncrasies of the particular data in question.

There is a long history of systematic sensitivity analysis (SSA) in CGE modeling that explores the sensitivity of results to changes in parameters.[24] This can be done in a variety of ways, but the use of SSA to understand external validity is most useful when the sensitivity analysis incorporates information on the precision of the estimates. That is, some parameters are precisely estimated with narrow confidence intervals while others may have quite large standard errors. Ignoring precision is especially problematic if very large elasticities come equipped with large standard errors. An import demand elasticity of 12 with a standard error of 6 might meet conventional statistical tests for significance, but the associated confidence interval on the trade response to a price shock is enormous.

Hertel *et al.* (2007) address this point in the context of a "Free Trade Area of the Americas" (FTAA) preferential trade agreement. They begin by estimating import demand elasticities following Equation (18.13), after first matching their data sample to the products and the country set involved in the subsequent trade liberalization experiment. Both point estimates and associated standard errors are provided, and several patterns emerge. Parameters for manufacturing sectors are estimated with much greater precision than those for mining and agriculture. This is likely because there are many more data points with greater price variation from which to identify parameters for trade in manufacturing. For example, machinery and equipment has a point estimate of 8.1

[24] See Pagan and Shannon (1987), Wigle (1991), Harrison and Vinod (1992), and Harrison *et al.* (1993).

with a standard error of 0.1, estimated from 44,386 data points, while natural gas has a point estimate of 34.4 and a standard error of 14.3, estimated from eight data points.[25]

Hertel *et al.* then use the precision of the estimates in their simulation of the effect of a FTAA. Essentially, they treat the elasticity parameters as draws from a distribution with means and standard deviations given by point estimates and standard errors of the parameters. By solving the model repeatedly, given these draws, they formulate confidence intervals around economic outcomes and welfare measures generated by the simulation. A key finding is that certain countries happen to trade goods for which parameter estimates lack precision. In these cases, there is significant variation in the simulated terms of trade and much wider confidence intervals on welfare. In cases where countries trade goods with tightly estimated parameters, welfare estimates also have narrow confidence intervals.

The external validity lesson for practitioners is 2-fold. Draw elasticities from studies that provide standard errors[26] and use them as part of sensitivity analysis. To be clear, however, parameter choice even for a parameter as critical as σ is just one source of uncertainty in modeled outcomes. Uncertainty about the underlying model structure itself is also important, even in the case where we have tight confidence intervals on parameter point estimates.

18.3.5.2.2 Cross-country differences: product composition and aggregation

A more serious problem for external validity is the possibility that the underlying parameters, σ_k from the utility function in Equation (18.4), are not the same for all countries and time periods. Leaving aside the possibility that underlying preferences are fundamentally different (which could be true, but is not especially helpful), there are two reasons why parameters might differ across country samples: product composition and the possibility that the true underlying model features a variable elasticity.

CGE models are typically limited by data and computational constraints to make use of relatively aggregated industries. Within each of these industries are many products that are quite distinct from one another and which may, therefore, feature different elasticities of demand. For example, a product category such as "Transportation Equipment" aggregates over cars, trucks, buses, motorcycles, associated parts for each, and much more. Broda and Weinstein (2006) estimate elasticities for product categories at different levels of aggregation (three-, four- and five-digits of the SITC classification, and up to 10

25 The underlying data are available at the six-digit level of the HS, but are aggregated up to (roughly) two-digit categories used in GTAP. The large difference in number of observations results primarily from the greater degree of disaggregation available in HS manufacturing trade categories — machinery and equipment are comprised of many hundreds of different HS6 product categories, each of which provides a separate price and quantity observation, while GAS has only one HS6 code. In addition, precisely because machinery and equipment is highly differentiated many countries export in this category, while relatively few export the homogenous product natural gas.

26 Broda and Weinstein (2006) and Broda *et al.* (2006) do not.

digits of the HS classification) and make these elasticities available to researchers via their website. This allows us to make comparisons of elasticities within each subsector. For example, within SITC 78 "Road Vehicles" there are 37 different product codes at the five-digit level. Leaving out two outliers with elasticities greater than 100, the simple average elasticity is 7.0 and the standard deviation across categories is 8.8.

What is the appropriate elasticity to use for a more aggregated sector? Let us assume that these point estimates are correct and, at the five-digit level, the same across importers, but that there are large differences across importers in the share of trade in each five-digit category. A country that primarily imports SITC 78520 "Bicycles" (estimated elasticity 1.6) faces very different elasticities than a country that primarily imports SITC 78319 "Public Transport Type Passenger Vehicles" (elasticity 13.2). As the composition of trade is, in fact, quite different across importers within an aggregate like "Transport equipment" estimates based on one country will be of dubious value when applied to another country.

18.3.5.2.3 Cross-country differences: variable elasticities

To this point we have treated the price elasticity of import demand as a constant and determined by a deep parameter in the utility function. This is consistent with the standard use of CES preferences and free entry monopolistic competition models of pricing. However, in alternative models of product differentiation the elasticity of import demand is not a constant. Instead, it reflects utility parameters, but also features of the underlying market structure such as the number of competing firms. To the extent that market structure differs across markets so will the elasticity of import demand.

Lancaster (1979) style models feature consumer preferences for an "ideal variety" located on a particular address in a finite product space. As product space is finite, entry causes firms to become closer substitutes and drive up the price elasticity of demand. In quadratic utility models such as Melitz and Ottaviano (2008), firms face linear demands with a finite choke price. Entry affects the intercept, the slope and the position on the demand curve, again raising the price elasticity of demand. In translog preference models such as Feenstra and Weinstein (2010), optimal pricing for a firm depend on that firm's market share. Again, entry drives up the price elasticity of demand.

Recently, these theoretical models of variable elasticities have found support in empirical work. Hummels and Lugovskyy (2009) use a generalized ideal variety formulation to show that the price elasticity of import demand depends on country size and *per capita* income, then confirm this prediction using cross sectional and panel data. Feenstra and Weinstein (2010) show that rising imports drive up the price elasticity of import demand and drive down markups.

Feenstra (1994) and the extension in Broda and Weinstein (2006) estimate import demand elasticities across multiple exporters for many products in US imports. Broda *et al*. (2006) use this methodology applied separately to the imports of 72 different

countries. They find very large differences across importers for a given SITC three-digit product. These elasticities are reported (sans standard errors) on the authors' website. For a given product, the coefficient of variation (standard deviation/mean) equals 2, meaning that an exporter with a demand elasticity one standard deviation above the mean has an elasticity twice as large as the mean. To put these numbers in perspective, the median elasticity (over all products and all importers) is 7.4, while the standard deviation across exporters is 15.5. It is not possible to discern exactly why these numbers are so different. Perhaps the standard errors associated with these estimates are enormous, which suggests a problem with the estimation technique. Perhaps it is an aggregation and product composition issue akin to the one just described. Or perhaps it reflects substantial differences across importers in the import demand elasticity for similar goods resulting from variable elasticity models and differences in market structure.

What is the external validity lesson for practitioners? It is tempting when viewing these results to suggest that modelers should incorporate endogenous entry and markups into their models. Short of that, it might be reasonable to treat preferences as CES, and import demand elasticities as constant, but allow the parameters to be specific to individual countries. This would recognize that variable elasticity models imply a different price elasticity of demand for the US than for Costa Rica, while treating any model-induced changes as small relative to those cross-sectional differences.

If this approach is taken, one should consider that many elasticity estimates in the literature are drawn from US imports and the US market may be in some sense the least typical market one might employ. An alternative is to turn to estimates that draw on the largest possible set of countries, as in Erkel-Rousse (2002), or allow elasticities to be market specific as in Broda *et al.* (2006).

18.4 EXPORT SUPPLY

As we indicate in Figure 18.1, the price and quantity of international trade flows depend jointly on demand- and supply-side behavior. Here we turn our attention to parameters that govern export supply. In many contexts export supply lacks a structural interpretation and is therefore not parameterized directly in structural CGE models. We review the contexts where either a reduced form or a structural interpretation is appropriate.

Our review indicates that parameterization of the supply side is much less visible and controversial than the literature on demand-side parameters. In our view a review of this literature is worthwhile for at least three reasons. (i) An important class of CGE models (i.e. single-country models) rely on such mechanisms to summarize trade behavior. (ii) Econometric estimates of demand-side parameters are confounded by supply-side behavior. (iii) An important new literature on firm heterogeneity provides a parsimonious structural framework for summarizing supply-side behavior and this is a feature that CGE modelers may wish to represent in future modeling efforts.

18.4.1 Mechanisms

In many CGE applications, there is no explicit parametric representation of export supply or even supply more generally. Producer behavior is captured by a cost function and a zero-profit condition. In this context there is no need to formally parameterize export supply. In multicountry models such as GTAP, demand-side parameters determine first-order trade responses, while supply-side responses are captured by general equilibrium effects. The cost function implies factor demands and factor market clearance equations determine factor prices. Given this, an increase in export prices eases the zero-profit condition, and the export sector expands until the zero-profit condition holds once more. A supply curve could be defined as the response of the cost function to increased output in a particular sector. However, rather than directly parameterizing the curve, its elasticity is implicitly defined by the general equilibrium structure of the model.

In single-country models, the rest of the world is not modeled directly, and this requires export supply behavior to be represented and parameterized explicitly in two places. (i) The modeler chooses a parameterization of import supply (i.e. the supply of exports by the rest of world to the single importer being modeled). This is usually treated as a reduced form schedule with its shape defined by a single elasticity parameter:

$$p_{\mathrm{M}}^{k} = \mu_{\mathrm{M}}^{k}\left(q_{\mathrm{M}}^{k}\right)^{\frac{1}{\omega^{k}}}, \tag{18.16}$$

where p_{M}^{k} is the import price in commodity k, μ_{M}^{k} is a scale parameter, q_{M}^{k} is the quantity of imports and $1/\omega^{k}$ is the inverse supply elasticity in commodity k.[27] Some models employ a small country assumption in which the importer is a price taker and faces an infinite elasticity of import supply. Others, such as the USAGE model at the US International Trade Commission (US ITC), for example, employs finite and positive elasticities of import supply to reflect the idea that the US is large enough to affect its import prices.

(ii) Modelers parameterize the supply of exports from the single country being modeled using a constant elasticity of transformation (CET) production technology that distinguishes goods by their destination market.[28] The simplest CGE models with CET technology have two production sectors — a domestic sector and an export sector — with an elasticity of transformation between the sectors of ω. Formally the technology takes the form:

$$L_i = \left(\alpha q_{\mathrm{d}}^{\frac{1+\omega}{\omega}} + (1-\alpha) q_{\mathrm{x}}^{\frac{1+\omega}{\omega}}\right)^{\frac{\omega}{1+\omega}}, \tag{18.17}$$

[27] One could account for multiple sources of imports by attaching origin subscripts to p_{M}^{k}, μ_{M}^{k}, q_{M}^{k} and possibly ω^{k}.

[28] Powell and Gruen (1968) introduce the constant elasticity of transformation technology. de Melo and Robinson (1989) develop general equilibrium implications for simple CGE models.

where L_i is a fixed stock of a factor production, α is a share parameter, and q_d and q_x represent the quantities the goods bound for domestic and export markets. Optimization subject to this technology generates an upward-sloping export supply function with a local elasticity of ω. It is straightforward to extend imperfect transformation between domestic sales and exports to multiple production sectors.[29]

To those unfamiliar with CET, this formulation is somewhat unintuitive. We more commonly model the limiting case in which the elasticity is infinite, the transformation schedule is linear and, for a given producer, goods have the same production cost regardless of where they are to be sold. Or, if there is a fixed cost associated with selling into a new market, costs are increasing in the number of markets to which firms sell. In contrast, the CET with a finite elasticity implies that firms minimize costs when they sell over a larger portfolio of destinations. Increasing the ratio of domestic sales to exports raises the opportunity costs of domestic sales.

How then to think about the intuition behind the CET? The most straightforward case is that, within sectors, the composition of export bundles differs substantially from that of goods intended for domestic consumption. As an example, consider our discussion in Section 18.3 regarding the disparate set of goods (trucks, buses, cars, bicycles and parts thereof) subsumed in "Transportation Equipment." These constituent goods within a sector likely differ in factor intensity of production. Even within narrower goods categories, export goods may require different technical standards or have other product characteristics that differentiate them in production from goods bound for domestic markets. This seems especially plausible in developing countries, which are frequently modeled in this framework. In these cases, we can think of the CET as representing concavity of the production surface arising from the usual comparative advantage reasons, but here it operates within, rather than across, sectors. As such, the CET framework is useful as a parsimonious way to represent imperfect production substitution between domestic supply and exports. However, because it is generally treated as a reduced form and not micro-founded it is difficult to know what the parameterization is actually capturing, and how it would respond to policy shocks.

Recent developments in the theory of heterogeneous firms in international trade provide a framework in which the nature of the tradeoffs between domestic and foreign sales is made explicit. In Melitz (2003) monopolistically competitive firms vary in their productivity, and face fixed costs of domestic production and of exporting. The most productive firms choose to sell domestically and to export; less productive firms sell only to domestic markets and the least productive firms exit. Most significantly for our

[29] A number of single-country models commissioned by the World Bank employ nested CET technologies to study options for trade liberalization that include preferential trading arrangements. Rutherford *et al.* (1993) study Morocco, employing an elasticity of transformation of 5 in the domestic–foreign nest, and 8 in the transformation between European Community and other markets. Eby-Konan and Maskus (1996) parameterize their model of Egypt in a similar way.

purposes, the composition of firms selling domestically differs from those engaged in exports. An upward-sloping export supply curve arises because the expansion of export sales occurs via the entry into exporting of marginally less productive firms charging higher prices. This is within sector heterogeneity, as hinted at in the preceding paragraph, but the nature of the heterogeneity is explicit.

Feenstra (2010) shows that the Melitz theory allows one to derive a CET between destination markets in which the export supply parameter is micro-founded and linked to structural parameters. In other words, the transformation function like that in (18.17) can be derived, except that the terms on the right-hand side are given a structural interpretation consistent with the Melitz technology. In the present context, Feenstra shows that ω can be linked to the structural parameters from Melitz as follows:

$$\omega \equiv \frac{a\sigma}{\sigma - 1} - 1, \tag{18.18}$$

where σ is the elasticity of substitution among varieties and a defines the shape of a Pareto distribution of firm-level productivities.[30] In a subsequent section, we will review empirical studies of the firm size distribution that put $\hat{a}$ between 3.5 and 5, and $\hat{\sigma}$ in the 3–4 range. These estimates imply reduced form CET parameterizations of roughly 3–7.

Recent evidence from the empirical literature suggests that there may be additional channels that affect firms' export supply responses. A growing literature has sought to respond to evidence presented by Kehoe and Ruhl (2009), which shows that trade responses to NAFTA were largest in goods with relatively low market penetration, *ex ante*. One of these explanations, put forward by Arkolakis (2010), is that exporting requires a marketing activity that features decreasing returns to scale. Highly efficient firms, with high levels of penetration into a given export market, are therefore less responsive to trade cost reductions than less efficient firms that have low levels of market penetration.[31]

18.4.2 Estimating reduced form schedules

The use of export supply schedules is most common in single-country models. In most cases, CGE modelers parameterize these schedules in what appears to be an *ad hoc* fashion, which is to say, without directly citing econometric estimates of the parameter.

[30] The quantity that responds with this elasticity is a variety-adjusted mass of firms that is relevant for welfare analysis. We shall represent Melitz quantities differently in subsequent section that describes structural estimation methods for the Melitz model.

[31] Eaton *et al.* (2011) summarize supporting evidence from the cross-section as follows: "(T)he general efficiency of a firm is very important in explaining its entry into different markets, but makes a smaller contribution to the variation in the sales of firms actually selling in a market."

Nonetheless, there have been efforts in the econometric literature to estimate such parameters and we review them here.

18.4.2.1 Single exporter

One common approach to estimating export supply relationships that approximates a CET form is to specify a revenue function representing GDP and to estimate a flexible functional form that links output in particular sectors to changes in relative prices.[32] The approach typically exploits time-series data, has a quite aggregated representation of sectors and focuses on the tradeoff between exports versus domestic sales. The production decision that is the focus of the analysis trades off amongst production sectors. The output of such papers is generally an export-supply elasticity, which can be used to parameterize a CET function in a policy model.

Kohli (1978), defines a translog restricted profit function in which representative firms in the production sector choose input demands to produce (negative) imports, exports, investment goods and consumption goods. The derivatives of this function define a system of supply and input demand equations that are estimated jointly, subject to homogeneity and symmetry restrictions. This system is estimated using annual time-series data for Canada from 1949 to 1972. Kohli's estimated export supply elasticity for Canada ranges from 1.5 to 2.2.

Several subsequent studies use this or related techniques to estimate rather low export supply and import demand elasticities. This suggests that the response of large aggregated sectors to annual relative price changes is rather sluggish. For example, Kohli (1993) applies a similar estimator to US data, finding export supply elasticities of magnitudes similar to Canada, and further that estimated export supply and import demand elasticities are falling over time. Diewert and Morrison (1986) estimate US export supply and import demand elasticities for the period 1967–1982. Export supply elasticity estimates are in the range of 0.33–0.38, suggesting an extraordinary resistance of quantities to changes in prices.

The estimation of GDP functions shares several key features with the estimation of Armington elasticities on the demand side. Sectors are usually quite aggregated, with variation in price and quantity indices driving the estimates. Much of the observed variation in prices is transitory, and the problem of simultaneity is ignored. Perhaps most puzzling is that these estimation techniques have typically been applied to large countries, where the identifying assumption of price-taking behavior is presumably weak.[33] The application of such estimates to policy models is of doubtful value.

[32] In situations where factor price information is available, rather than factor quantities, a profit function is estimated rather than a revenue function.

[33] The main reason that the countries used are large is likely the limitations of available data. One needs a long time series of detailed information of prices and quantities of several goods and factors. Since this estimation was considerably more common several decades ago, there would have been few countries with such data available.

18.4.2.2 Single importer

Many single-country CGE applications feature a reduced form elasticity of import supply (that is, export supply from the rest of the world). In trade policy applications it is an important parameter because this elasticity determines the optimal level of the tariff. Since price and quantity of import data are available, this elasticity can, in principle, be estimated from the perspective of a single importer. However, our review of the literature suggests that these parameters are not frequently estimated nor does it seem that the estimates that do appear are frequently implemented in the CGE literature.[34]

One example we could find of a direct reliance on econometric estimates is the initial US ITC model for the US that parameterizes the inverse import supply elasticity using econometric estimates from Haynes and Stone (1983).[35] These authors propose an empirical model for export supply aggregated over all major trading partners to yield a single quantity of imports in each year for the US:

$$PM_t = \beta_0 + \sum_{j=0}^{4} \beta_{1j} QM_{t-j} + \sum_{k=1}^{2} \beta_{2k} WP_{t-k} + \beta_3 TY_t + \beta_4 CY_t + \mu_t. \qquad (18.19)$$

The aggregate import price index (PM_t) depends on contemporary values of the quantity index (QM_t), four quarterly lags of the quantity indices, an aggregate of wholesale price indices among major US import sources (WP_t) and two quarterly lags — trend income in the source country (TY_t) and capacity utilization (CY_t) in the supplying country. Following the IV strategy of Fair (1970), contemporary values of all the exogenous variables and lagged values of the endogenous values are used to instrument for quantities. The contemporaneous coefficient on quantities is 0.13, so the import supply elasticity used in the US ITC model is $1/0.13 = 7.7$.[36]

The estimation methods employed here suffer from many of the same weaknesses as the time-series methods discussed earlier in relation to demand parameters. The identification comes from temporary shocks, even though most CGE experiments will consider permanent shocks. The use of unit value indices in estimating price quantity relationships is likely to be biased toward zero if quantities are measured poorly. While IV is used, it is not clear to us why these are valid instruments (i.e. why they are correlated with demand related shocks to prices but uncorrelated with the errors in the supply

34 When finite parameterizations of import supply are used, they are often appear to be chosen in an ad hoc fashion. Good practice seems to be sensitivity analysis over a wide set of implied rest of the world export supply elasticities. For example, in an investigation of trade policy consequences of trade restrictions in cheese and sugar products, US ITC (2002) specify product-level import supply elasticities for these commodities of approximately 7, but conduct sensitivity analyses in which all import supply elasticities are set to 50, which approximates price-taking behavior.

35 See US ITC (1989, p. D-12).

36 Note that this is a contemporaneous and not a long-run effect. Long-run estimates are calculated in Haynes and Stone, but not employed in the US ITC model. The econometric model predates, and so ignores, cointegration, which has been shown by Gallaway *et al.* (2003) to be important in demand-side estimation.

equation). All this points to coefficients biased toward zero. However, in this context, where models use the inverse supply elasticity, this bias causes authors to assume high degrees of substitutability between home and foreign destinations.

18.4.3 Supply–demand systems

In Section 18.3 we described in detail an estimation strategy devised by Feenstra (1994) that jointly estimates export supply and import demand elasticities using import data. Here we focus on export supply parameters. The estimation strategy exploits a panel of annual import flows by a single importer, disaggregated by the exporting source country and detailed commodity. A critical identifying assumption is that for a given importer the elasticities of import demand and export supply are the same for all exporters.

While Feenstra reports estimates for a small set of commodities, Broda *et al.* (2008) estimate export supply elasticities facing 15 importers in up to 1200 HS four-digit products and provide data on over 12,000 elasticities on line.[37] The magnitude of these elasticities can be used to measure the extent of an importer's terms-of-trade power. The median value is 1.59, which implies that a 1% decrease in quantities imported lowers the export supply price by 1.59%. This is a significantly higher elasticity than is found in the reduced form literature and much higher than is typically used in the modeling area.[38]

This is far the most comprehensive set of econometrically estimated export supply elasticities available to the CGE modeling community. Should they be trusted and used? Consider first some concerns with the underlying assumptions and the data employed. The identifying assumptions are the same as Feenstra (1994), but because Broda *et al.* apply the estimator to many different importers, the authors can estimate different export supply elasticities facing each importer. This is somewhat curious: it assumes that Italy and China have the same export supply elasticity when selling textiles and apparel to the US (which implies they face the same opportunity costs of production), but that China may have a different export supply elasticity when selling to the US and to Japan.

Using the Broda *et al.* data, we can evaluate the assumption of equal export supply elasticities in two ways. First we use an analysis of variance to ask: how much of the total variance in elasticities is specific to each product? If elasticities are the same across countries for a given product, the answer should be an R^2 of 1. Instead the ANOVA shows the adjusted R^2 is 0.04, i.e. 96% of the variation is within, rather than across, product lines. We can also examine how much elasticities vary across countries for a given product by calculating the coefficient of variation. For the median product the coefficient of variation is 2.24, meaning that the country one standard deviation above

[37] See: http://www.columbia.edu/~dew35/TradeElasticities/TradeElasticities.html.

[38] The median actually understates the magnitude of these elasticities because the estimates are highly skewed. This can be seen by examining percentiles of the elasticity distribution. The 25th percentile elasticity is 0.5, the 75th percentile is 11.6 and the 95th percentile is 752. Given this skew, the mean elasticity is 85.

the mean has an elasticity 2.24 times greater than the mean country. We find it difficult to square these very large differences across countries with the underlying identifying assumption that trade elasticities are common.

A second concern is the magnitude of the elasticities and the large degree of pricing power they assert. This may be an artifact of the time horizon — the shocks to prices are based on annual data and so quantity responses may primarily capture short run responses. Results might be quite different for experiments that model long-run policy changes.

Finally, Soderbery (2010) dissects Feenstra's estimation strategy in a Monte Carlo exercise. He shows that while the Feenstra method is asymptotically efficient, the relatively short panels in available international trade data are potentially a source of bias. In Monte Carlo simulation data that replicates key features of the Feenstra data, he finds that while the demand parameters can converge quite quickly to the true parameters, the supply parameters are severely biased in small samples. This may be especially relevant to Broda *et al.* given the short (9-year) length of their data panel. Since these publicly available estimates do not contain standard errors or confidence intervals, they offer little guidance for sensitivity analysis around the parameters of interest.

With all that said, the elasticity of export supply facing an importer is an important parameter, especially in single-country CGE modeling. It is a critical determinant of the optimal tariff rate and importers facing a finite export supply elasticity can use trade policy to affect its terms of trade. Indeed, the relationship between export supply and optimal tariff setting is, in fact, the main focus of Broda *et al.* After estimating export supply elasticities using the Feenstra (1994) methodology, they show that across-product variation in tariffs for 15 non-WTO member countries is positively correlated with inverse supply elasticities. That is, tariffs are highest when countries face a less elastic export supply curve and so have the greatest terms-of-trade power. They also find that inverse supply elasticities are positively correlated with measures of US protection (such as antidumping investigations), which may not have been affected by US membership in the WTO.

Despite the econometric concerns raised above, the inverse supply elasticities estimated by Broda *et al.* appear to contain useful information about the degree of market power held by individual countries. Since this is the precise context in which these elasticities matter for single-country CGE models, modelers should display keen interest in the developments in this literature going forward.

18.4.4 Supply-side reinterpretation of bilateral trade responses

In Section 18.3 we described a literature that identifies import demand elasticities by relating variation in bilateral trade to variation in trade costs. Standard trade models such as Armington and Krugman can be manipulated to produce a bilateral trade prediction like Equation (18.12). In these models key variables such as prices and the number of

firms exporting are specific to an importer or an exporter, but not a bilateral pair. This allows the econometrician to use a specification like Equation (18.13) that holds constant export supply conditions using fixed effects or differencing strategies, and then uses bilaterally varying trade costs to identify a structural demand parameter σ.

Recent theoretical developments due to Eaton and Kortum (2002), Melitz (2003), and Chaney (2008) suggest, however, that the empirically estimated responses can just as easily be given a supply-side interpretation. We focus our discussion on the structure in Melitz and Chaney; the insights in Eaton and Kortum (2002) are similar in spirit.[39]

Recall from our discussion in Section 18.4.1 that Melitz (2003) features monopolistically competitive firms that vary in their productivity, and face fixed costs of domestic production and of exporting. The most productive firms choose to sell domestically and to export; less productive firms sell only to domestic markets, and the least productive firms exit. In this environment, Chaney (2008) draws out the implications for bilateral trade when trade costs change.

In a model with homogeneous productivity, a fall in trade costs induces a rise in bilateral trade, with an elasticity of σ. Chaney shows that, with heterogeneous firms, σ has two counteracting effects on bilateral trade. The intensive margin is the same as Equation (18.12): holding constant the number of firms shipping, a 1% reduction in trade costs corresponds to a 1% reduction in delivered prices and the value of bilateral shipments rises in proportion to the price elasticity of demand, $(\sigma - 1)\%$.[40]

However, the Melitz model also introduces an extensive margin, defined as export sales by firms not previously in the market. The fall in trade costs allows less productive firms to enter the export market; how much these new firms export compared to the incumbents depends on how much less productive they are. Assuming a Pareto distribution for productivity, the elasticity of the extensive margin of trade with respect to trade costs is $a - (\sigma - 1)$.

The intuition for the extensive margin response is as follows. The Pareto parameter a describes the degree of heterogeneity in productivity. When "a" is large, firms are similar in productivity and so new exporters offer goods at prices only slightly higher than incumbents. When "a" is small, firms are highly heterogeneous and so new firms offer goods at much higher prices than incumbents. How these price differences translate into export sales depends on the demand parameter. The extensive margin, export sales by new firms, is largest when new firms sell at prices similar to incumbents (high a) and consumers are less sensitive to those price differences (low σ).

[39] Melitz focuses on within-industry heterogeneity in firm productivity while Eaton and Kortum examine heterogeneity in productivity at the commodity level. Conceptually this is not very different since the underlying demand system is the same, and Melitz's "industry" is in fact all of the economy except for a numéraire.

[40] This is the same elasticity as in Equation (18.12), except that we are now dealing in c.i.f. (cost, insurance and freight) valuations.

Chaney shows that if we combine the two effects, the overall elasticity of trade to trade costs ζ depends only on the productivity distribution, defined by the parameter a:

$$\zeta = -\frac{d\ln X_{ij}}{d\ln \tau_{ij}} = \underbrace{(\sigma - 1)}_{\text{intensive margin}} + \underbrace{(a - (\sigma - 1))}_{\text{extensive margin}} = a. \tag{18.20}$$

Higher values for the demand parameter σ makes the intensive margin more responsive to trade costs (small tariff cuts generate large increases in trade for incumbent firms), but it makes the extensive margin less responsive (new firms entering the market sell little). In this stylized model, trade volumes are, on net, independent of the demand parameter σ.

It is useful at this point to return to our discussion on estimating import demand elasticities. One way to view the Melitz/Chaney interpretation of bilateral trade is that regressions like those run by Hummels (2001), Hertel *et al.* (2007) and Romalis (2007) do produce a meaningful reduced form measure of the trade response to changes in trade costs. However, their structural interpretation of that parameter is wrong, reflecting heterogeneity in firm supply rather than the price elasticity of demand. The same critique may apply to Feenstra (1994), Broda and Weinstein (2006), and Erkel-Rousse and Mirza (2003), and indeed to the time-series literature on import price elasticity to the extent that their price variation is induced by shocks to trade costs or exchange rates.

To understand this point econometrically, note that these papers employ a crucial identifying assumption. If the number of varieties and prices from the import demand Equation (18.11) do not vary over bilateral partners, we can use firm fixed effects or differencing strategies to eliminate them and isolate just the intensive margin of adjustment. Melitz/Chaney can be interpreted as saying that these variables are bilaterally varying and so not differenced out, and moreover, are correlated with trade costs because of entry/exit of firms into exporting. Thus, the structural interpretation of the estimated trade response remains unidentified in the cross-section. Given a particular specification of the regression, the response of bilateral trade to variation in trade costs can be attributed to either a supply or a demand parameter.[41]

A similar point arises if we consider other sources of firm heterogeneity. For example, suppose firms differ in the intermediate inputs they purchase. This creates a tendency for firms to colocate with their input suppliers in order to avoid trade costs for back-and-forth shipment. Some prominent examples of this mechanism in varying theoretical contexts can be found in Krugman and Venables (1986), Rossi-Hansberg (2005),

[41] Arkolakis *et al.* (2010) show that for some applications it is not necessary to separately identify structural supply and demand response parameters. The models considered are relatively simple trade models (iceberg trade costs, imperfectly elastic factor supply, and a constant trade response across origin and destination pairs). Furthermore, Arkolakis *et al.*'s focus is limited to welfare measurement. CGE models are typically interested in welfare measurement, but also a broader set of counterfactual outcomes including industry-level output and factor returns.

Baldwin and Venables (2010), and Yi (2010). The implication in the current context is that the number of firms exporting from a location is not a fixed constant, but varies bilaterally depending on the product composition of production in a destination region. Evidence for this claim is found in Hillberry and Hummels (2008), who rely on zip-code-level data on intra-national shipments within the US. They show that the effect of distance on trade is highly nonlinear, with the volume of shipments falling rapidly over the first 100 miles. The vast majority of this response comes through changes in the number of firms shipping to proximate versus distant locations, not through changes in the average value of shipments per firm.

What do models of firm heterogeneity imply for the estimates of import demand discussed in Section 18.3? The key question is whether the extensive margin in (18.20) is active at the level of aggregation and in the time horizons employed by authors who seek to identify demand parameters. It may be that the extensive margin is more active at longer time horizons, i.e. firms are better able over the long run to enter and exit in response to shocks in trade costs. If entry and exit is slow, the extensive margin effect may be less important at higher frequencies.

Related, some nuance is required to map the models into the associated parameters estimated given the level of aggregation at which each is concerned. Melitz (2003) describes a distribution of productivity drawn for all firms economy-wide and is silent on whether these firms exist within or across a particular category of traded products. Meanwhile, many empirical studies of import demand elasticities examine changes in trade within product categories, in some cases using exceedingly narrow product categories such as 10-digit HS groups for Feenstra (1994) and extensions in Broda and Weinstein (2006). The extent to which the Melitz/Chaney interpretation captured in Equation (18.20) correctly describes which parameters are being identified then depends on where and when the extensive margin is active.

Put another way, when Chaney says that the extensive margin cancels out the effect of the demand parameter operating through the intensive margin, how are we meant to read that? Is it a statement about how trade costs affect aggregated bilateral trade or trade at the industry or at product-line level? If the model addresses itself to larger industry or economy-wide aggregations, then the extensive margin combines entry into trade of firms within a product classification as well as entry of product classifications themselves.[42] Further, if the main entry action is happening across product codes, rather than within, the more the econometrician disaggregates the more they are isolating an intensive margin (and therefore demand parameters). For example, imagine that every

[42] Besedes and Prusa (2006a, 2006b) work with trade data that is disaggregated across narrow product categories, but not disaggregated at the firm level. They show that entry/exit of entire industries, as opposed to entry/exit of firms within an industry, is an important adjustment margin for trade even at high frequencies. Helpman *et al.* (2008) study how selection into trade on the extensive margin affects estimated effects of trade costs on aggregate trade flows.

firm in a Melitz setting corresponds to a single one of the 5000 distinct HS six-digit codes available in modern trade data. Then if the econometrician performs all estimates within a given HS6 code, they are identifying intensive margins and therefore demand parameters. If the econometrician begins to aggregate these product lines into larger and larger industrial aggregates, the role of the extensive margin looms larger. This is a possible explanation for why estimated trade elasticities are sensitive to level of aggregation, with estimated magnitudes dropping the more aggregated are the sets of goods employed.

One possible way to distinguish demand and supply parameters is to identify independent estimates of the heterogeneity parameter a. This can be done by estimating the empirical distribution of firm sales. If firm productivity levels are Pareto distributed, and the elasticity of substitution between firm output is σ, as in Melitz, the distribution of firm sales should be distributed according to the Pareto distribution. The shape parameter that governs this distribution, ζ, has a structural interpretation:

$$\zeta = a/(\sigma - 1).$$

The ζ parameter has been estimated empirically, using firm-level data, by authors such as Axtell (2001) and di Giovanni *et al.* (2011). Several other authors have attempted to separately identify the a and σ parameters by (i) exploiting cross-country variation in the firm size distribution and (ii) imposing further economic structure on the estimating system. We review these efforts in Section 18.5.

18.5 GRAVITY, TRADE COSTS AND STRUCTURAL ESTIMATION

In the preceding sections we discuss econometric approaches to identifying trade elasticities. Recent studies have blended econometric techniques with other approaches including calibration and structural estimation to come at fundamentally similar problems from different angles. Some relate variation in trade flows to variation in trade costs to identify the unknown trade elasticity σ. This approach assumes that we know trade costs fully, or at least that unmeasured costs are uncorrelated with included costs. Other approaches focus on identifying unknown trade costs themselves; these take trade flows and σ as given and extract trade costs. Still others combine trade flows with general equilibrium theories of the determination of prices to structurally identify parameters in the supply side of the economy. In this section we review this literature in order to demonstrate the relationship between these seemingly disparate approaches, to understand which sorts of approaches are useful in different contexts, and finally to draw out the merits of structural estimation methods, and their relation to linear econometric methods.

Consider a version of Equation (18.12), describing the value of shipments from i to j:

$$\ln X_{ij} = \sigma \ln \alpha_i - \sigma \ln p_i\left(1 + \tau_{ij}\right) + \sigma \ln P_j + E_j + e_{ij}. \tag{18.21}$$

We generically represent supply characteristics in exporter i using α_i and add an error term. The linear approach to estimating σ from variation in trade costs requires the econometrician to eliminate importer- or exporter-specific terms using either origin and destination fixed effects, or differencing the data relative to other exporters or importers. This approach treats the structural interpretation of the origin- or destination-specific terms as unimportant. In other contexts we may not be able to fully measure trade costs, and the origin and destination terms may be of direct interest. This requires a general equilibrium approach in which prices and supply characteristics are not treated as nuisance parameters but instead defined (or imputed) as part of the model outcome.

18.5.1 Approaches to identification

In this section we consider various approaches to the identification of the structural parameters in (18.21). A common approach is to take an estimate of σ from external sources and impose it in order to achieve estimation. Other approaches bring in additional data — either information on measured trade costs, or information that informs p_i. Various approaches also impose additional assumptions that limit the parameter space.

18.5.1.1 Methods that impose an outside estimate of σ

Suppose we know trade flows and the elasticity of import demand, but do not know the full characteristics of trade costs afflicting trade. We can use the structure of Equation (18.21) to extract these trade costs. Note that all variables in Equation (18.21) have either an i or a j subscript. By expressing trade flows in differences we can eliminate all these variables, leaving only trade costs. The most stark example of this approach is in Chen and Novy (2011) and Jacks *et al.* (2011). They write Equation (18.21) in levels and benchmark bilateral trade against domestic shipments. That is, they compare i's exports to j relative to i's sales to its own domestic market:

$$X_{ij}/X_{ii} = \frac{E_j\left(1+\tau_{ij}\right)^{-\sigma}P_j^{\sigma}}{E_i(1+\tau_{ii})^{-\sigma}P_i^{\sigma}}. \tag{18.22}$$

This eliminates all variables specific to the supply side, leaving only demand-side characteristics and trade costs. A similar expression for j's sales X_{ji}/X_{jj} contains the same expenditures and price terms from (18.22); writing this as a double difference $(X_{ij}/X_{ii})/(X_{ji}/X_{jj})$ eliminates all remaining variation except for trade costs and σ. With some manipulation they can extract an average measure of bilateral trade costs from trade flows by imposing an outside estimate of σ:

$$\kappa_{ij} = \left(\frac{\tau_{ij}\tau_{ji}}{\tau_{ii}\tau_{jj}}\right)^{\frac{1}{2}} - 1 = \left(\frac{X_{ij}X_{ji}}{X_{ii}X_{jj}}\right)^{\frac{1}{2(\sigma-1)}} - 1. \tag{18.23}$$

This approach is conceptually very similar to the econometric approach, in that it eliminates rather than models variables with *i*- or *j*-specific variation. Rather than estimating the response of trade to explicitly measured costs such as tariffs or transportation costs, it leans heavily on the structure of the model to infer them. It assumes there are no error terms and that all remaining variation in bilateral trade is due to trade costs. Given these strong assumptions and knowledge of a single parameter, one can proceed to calculate trade costs implied by the level of trade.

Against this backdrop, consider the traditional method of calibrating CGE models. Trade costs are measured explicitly, though incompletely. The general equilibrium model includes structural representations of prices and expenditures.[43] The focus is usually on exact calibration to the fitted flows, so the error term $e_{ij} = 0$ by assumption. In order to fit the trade data X_{ij} exactly, the calibrator has two free parameters, σ and a model residual α_{ij}, which is made more general than the form in (18.21) by adding *ij* subscripts. This can be interpreted as "tastes" that are specific to bilateral pairs, or as unmeasured trade costs like Chen and Novy (2011) and Jacks *et al.* (2011). As there is one more parameter to be fit than there are pieces of available data, the structural model is under-identified, but if we impose a value of σ, this completes the model calibration.

How different is the calibration exercise in a typical CGE model than simply attributing all residual variation to trade costs? One way to think about this is to ask how much of the variation in observed import shares is explained by the residual parameter α_{ij}. Hillberry *et al.* (2005) investigate this using a "calibration-as-estimation" experiment using GTAP. Holding fixed region *j*'s total expenditures on each commodity, variation in import shares is driven by explicitly measured bilateral trade costs (tariffs and transport costs) and variation across *i*'s in the residual parameter α_{ij}. For each GTAP sector, Hillberry *et al.* conduct a counterfactual exercise in which they reduce all measured bilateral trade costs to zero, then calculate the effect this has on the bilateral variation in import shares. If α_{ij} is relatively unimportant, removing all measured bilateral trade costs should significantly alter the bilateral distribution of trade. In 33 of the 46 sectors, they calculate that less than 20% of the variation in bilateral import shares disappears once all trade costs are removed. This suggests that the α_{ij} residuals are responsible for the large majority of bilateral variation in the calibrated data. Hillberry *et al.* also conduct sector-by-sector exercises in which they calculate the value of σ that is necessary to attribute all of the variation in import shares to trade cost measures in GTAP. The fitted values of σ are generally much larger than the default GTAP values of σ.

The interplay between parameterizing taste parameters, trade costs, and σ also plays out in the literature that investigates border costs and their welfare consequences. This

[43] Expenditures are normally data, but there are equilibrium conditions linking income and expenditure, as well as equations that link income to the parameters of the model. Prices depend upon the structural parameters via the equilibrium conditions of the model.

literature began with McCallum (1995), who used reduced form regressions to show that intra-Canadian trade flows greatly exceeded US−Canada trade flows, controlling for incomes and distance between partner regions. The initial estimates were all described in terms-of-trade value responses. To get at the *ad valorem* equivalent of these costs, and to make welfare claims about these costs, requires information on σ.

Anderson and van Wincoop (2003) address the border puzzle by developing an estimation procedure that clarifies the structural interpretation of measured trade responses. A key point is that narrowly focusing on the slope of the import demand curve misses important non-linear effects operating through the price index P_j. Their primary innovation relative to linear estimation is to introduce an equation that determines these price indices as a function of trade costs in general equilibrium.[44] Another contribution of the paper is to conduct counterfactual analyses using structural inferences about trade costs. Counterfactual analyses still requires a choice of σ in order to infer the *ad valorem* equivalent of distance and border costs and to govern counterfactual trade responses.

Balistreri and Hillberry (2007) revisit the estimation in Anderson and van Wincoop (2003). This is primarily a methodological piece that contributes by introducing a more general approach to structural estimation, imposing full theoretic consistency on the estimation model, and conducting general equilibrium counterfactual analyses.[45] The paper also considers the implications of pooling US and Canadian data sources. The paper's contribution in terms of the structural parameterization of the model is to relax an assumption of symmetric unmeasured border costs and to bound the estimates under the assumption that border costs are always non-negative. This procedure involves introducing an additional parameter to capture border cost asymmetry and then restricting the tariff-equivalent of the border cost to zero in each direction in order to produce bounds on the estimates.

The progression in the border cost literature is instructive. It begins with purely reduced forms, with trade responses later interpreted through the lens of a model like Equation (18.21). It ends up with increasingly formal structural approaches that examine not just the response of the import demand curve, but the effects of these costs on the full general equilibrium.

18.5.1.2 Methods that exploit price data to infer τ and estimate σ

Equation (18.21) is written as a demand-side equation, where the measured trade elasticity is identified as the demand-side parameter σ. As we note in Section 18.4, there are also theories in which the response of bilateral trade flows to variation in

[44] They assume an endowment economy so that prices are determined by trade cost-induced shifts of the demand curve, and price indices, which they describe as "multilateral resistance terms" can be solved from the system of economy-wide demands.

[45] For example, relative to the theoretical benchmark, Anderson and van Wincoop introduce a regression constant as a free parameter when the constant has a structural interpretation in the theory.

trade costs can be interpreted structurally as a supply-side parameter. Eaton and Kortum (2002) derive a parsimonious Ricardian theory of trade that summarizes the trade response in terms of a single structural parameter. This parameter defines the shape of a Frechet distribution that describes the evolution of comparative advantage across products.

Eaton and Kortum propose an estimation technique for identifying the Frechet parameter. To compare the identification strategy to that in Equation (18.21), the method supplements the trade flow information with data on retail prices p_i for 50 manufactured goods. They assume that price differences for homogeneous goods in two locations must not exceed bilateral trade costs, because an arbitrage condition bounds the price differences. Calculating the ratio of retail prices for the same good in two locations, $\tau_{ij} \geq p_j/p_i$, allows them to infer the magnitude of bilateral trade costs. Eaton and Kortum devise a simple method of moments estimator that relates bilateral to domestic trade and the implied bilateral trade cost. This pins down the supply-side structural trade response parameter, which they estimate to be 8.28.

Simonovska and Waugh (2011) develop a simulated method of moments estimator to address a perceived shortcoming in the Eaton and Kortum estimation. The theory implies a continuum of goods, and corresponding price gaps for each good in the continuum. As Eaton and Kortum infer trade costs from the tail of an empirical distribution defined by a small number of prices, rather than from an infinite number of prices as implied by a continuum, Simonovska and Waugh argue that Eaton and Kortum's approach systematically understates bilateral trade costs. The Simonovska and Waugh procedure treats the price draws as a sample from much larger distribution, and calculates the trade response parameter consistent with the larger implied trade costs. Their estimate implies a much smaller aggregate trade response, in the neighborhood of 4 rather than 8.

The merits of these two exercises depend upon one's willingness to accept the translation of retail price gaps as concrete measures of bilateral trade costs. On the one hand, retail price gaps may incorporate a more comprehensive set of costs than is typically contemplated in studies that make use of a directly measured set of trade barriers such as transportation costs and tariffs. On the other hand, price gaps include retail and wholesale markups that are unlikely to be arbitraged away via international trade. The inferences drawn about trade responses depend on one's assumptions about the interaction of the retail trade sector with international arbitrage conditions.

18.5.2 Structural estimation of theories with firm heterogeneity

In the monopolistic competition derivation of (18.21) that follows Krugman (1980), $X_{ij} = n_i p_i q_{ij}$. The variables n_i and p_i have no j subscript because all firms in i sell to all

markets and all prices in i are set with the same markup over identical marginal costs. This implies that all of the observed ij variation in trade flows is attributed to traded quantities, which are determined by a demand equation.

However, a considerable body of evidence suggests that the number of exporting firms does vary bilaterally, with more productive firms serving more markets. Recent theories of bilateral trade, notably Melitz (2003), offer an explanation for this evidence, and introduce a new structural interpretation of the bilateral trade responses. In Section 18.4 we reviewed the Chaney (2008) argument that trade responses should be interpreted in terms of supply-side responses. Here we review approaches to estimating the key supply parameter in Melitz (2003), along with other structural parameters relevant to firm heterogeneity.

We begin with a structural estimation/calibration approach proposed by Balistreri *et al.* (2011). These authors adapt an approach to structural estimation proposed by Balistreri and Hillberry (2007) to the parameterization of the Melitz model. The econometric model minimizes the differences between observed and fitted flows, subject to the equilibrium conditions defined by the Melitz (2003) model.[46] Balistreri and Hillberry (2008) argue that estimation along these lines provides a transparent mapping between the estimating model and the numerical general equilibrium used in counterfactual analysis.

Our purpose here is to compare the procedures in Balistreri *et al.* to other approaches reviewed above. The equilibrium conditions derived in Balistreri *et al.*, together with an equilibrium condition in Chaney (2008), can be manipulated to produce a bilateral trade prediction like that in (18.21). Specifically, we can relate the model consistent fitted flows $\hat{X}_{ij}$ to equilibrium outcomes such that $\hat{X}_{ij} = n_{ij}\tilde{p}_{ij}\tilde{q}_{ij}$, where the ~ notation over p and q indicates average values along each bilateral route.

As in Melitz (2003) and Chaney (2008) firms are monopolistically competitive, face CES preferences, and draw productivity from a Pareto distribution. In the presence of fixed and marginal trade costs, the most productive firms charging the lowest prices enter export markets. Balistreri *et al.* link the equilibrium productivity level of the marginal firm operating on the ij route φ^*_{ij}, to the proportion of firms that are active along that route:

$$\varphi^*_{ij} = b \Big/ \left(\frac{n_{ij}}{m_i}\right)^{\frac{1}{a}}. \tag{18.24}$$

The form of (18.24) follows the cumulative density function of the Pareto distribution, with m_i representing the equilibrium number of firms in i taking a productivity

[46] The econometric problem can be defined as Mathematical Program with Equilibrium Constraints (MPEC). Su and Judd (2011) develop MPECs as a general approach to estimation of structural estimation models.

draw and n_{ij}/m_i represents the proportion that are active along a given ij route. Chaney (2008) expresses φ^*_{ij} as:

$$\varphi^*_{ij} = \frac{\sigma^{\sigma-1}}{\sigma-1}\left(\frac{f_{ij}}{E_j}\right)^{\frac{1}{\sigma-1}}\frac{w_i\tau_{ij}}{P_j}, \tag{18.25}$$

where f_{ij} represents the bilateral fixed costs of serving market j from i, E_j destination market expenditure (GDP), w_i wages in region i, τ_{ij} are iceberg *ad valorem* trade costs between i and j, and P_j a CES price index of all varieties sold in j. Combining (18.24) and (18.25) allows the number of firms shipping to a destination n_{ij} to be expressed in terms of standard gravity variables:

$$n_{ij} = \gamma\left(\frac{f_{ij}}{E_j}\right)^{\frac{a}{\sigma-1}}\left(\frac{w_i\tau_{ij}}{P_j}\right)^{-a}, \tag{18.26}$$

where γ represents a collection of structural parameters. Note that the elasticity of the number of firms entering a market to bilateral trade costs is $-a$. This is the extensive margin prediction derived in Chaney, expressed in another way.

It is then possible to express average productivity, prices, and quantities along each bilateral routes by linking each to the behavior of the marginal firm:

$$\tilde{\varphi}_{ij} = \left(\frac{a}{a+1-\sigma}\right)^{\frac{1}{\sigma-1}}\varphi^*_{ij} \tag{18.27}$$

$$\tilde{p}_{ij} = \frac{\sigma}{\sigma-1}\frac{w_i\tau_{ij}}{\tilde{\varphi}_{ij}} \tag{18.28}$$

$$\tilde{q}_{ij} = \frac{\tilde{p}_{ij}{}^{-\sigma}E_j}{P_j^{1-\sigma}}. \tag{18.29}$$

Combining (18.25), (18.27), (18.28) and (18.29) produces an expression for average firm revenues:

$$\tilde{p}_{ij}\tilde{q}_{ij} = \frac{a\sigma}{a+1-\sigma}f_{ij}. \tag{18.30}$$

Interestingly, all of the standard gravity variables are eliminated from this expression, with average revenues expressed solely in terms of structural parameters and bilateral fixed costs. The intuition for (18.30) is that marginal firm revenues imply a constant mark-up over fixed costs, and the Pareto distribution links average firm revenues to marginal firm revenues via another constant mark-up. One can thus understand (18.30) as the zero cut-off profit condition for the ij route, expressed in terms of average firm revenues.

Using $\hat{X}_{ij} = n_{ij}\tilde{p}_{ij}\tilde{q}_{ij}$ one can see that the product of Equations (18.26) and (18.30) generates a predicted fitted value for bilateral trade. Expressing this in terms similar to (18.21):

$$\ln\hat{X}_{ij} = \ln\frac{\gamma a\sigma}{\sigma-1} - a\ln w_i + \ln m_i - a\ln\tau_{ij} - \frac{a+1-\sigma}{\sigma-1}\ln f_{ij} - a\ln P_j + \frac{a}{\sigma-1}\ln E_j. \tag{18.31}$$

As in (18.21), we have a bilateral trade prediction in terms of the model variables and the structural parameters. There are now additional variables, notably f_{ij}, which are potentially unobservable to the econometrician.[47] Linear regression is unable to proceed in this instance, but structural estimation may be able to jointly calculate f_{ij} and other parameters of the model under certain assumptions. Balistreri *et al.* illustrate one method for doing so that is closely related to the discussion earlier in this section.

The model remains under-identified without further assumptions (notably imposing σ, as well as normalizations of some of the other structural parameters). Balistreri *et al.*'s preferred estimate also imposes one additional parameter on the estimation, the elasticity of trade costs with respect to distance, δ. This approach is familiar from the literature on trade costs, distance and border effects. Quantities of trade flows are highly correlated with trade cost proxies such as distance. Translating the effect of distance on trade quantities into *ad valorem* equivalents requires both σ and δ.

Conditional on these assumptions, a procedure that minimizes:

$$z = \sum_{ij}\left(\ln X_{ij} - \ln\hat{X}_{ij}\right)^2,$$

can complete the calibration of the general equilibrium system.[48] This produces an estimate of the key supply-side parameter $\hat{a} = 4.6$, which is largely consistent with what is observed in the firm-level evidence we discuss shortly.

18.5.3 Evidence from firm-level data

It may seem surprising that one can extract a parameter describing firm heterogeneity exclusively from aggregate data on bilateral trade. This is only possible because the Melitz model with Pareto distributed efficiency is so parsimonious. Note that aggregate trade flows combine two independent and potentially measurable variables n_{ij} and $\tilde{p}_{ij}\tilde{q}_{ij}$. In cases where data on these two variables are available, the predictions of

[47] m_{ij} is also unobservable, but is linked quite closely to the factor supply in region i. The factor supply might be considered an observable characteristic.

[48] Note that this is a calibration to the fitted flows $\hat{X}_{ij}$, not exact calibration as is normally done in CGE models. Exact calibration requires a structural interpretation of the error term and Balistreri *et al.* apply one interpretation that treats residuals as attributable to idiosyncratic bilateral fixed costs.

the model can be tested and/or a richer parametric structure estimated. When firm-level data is available, one can also estimate other properties of the firm-level sales distribution, not just average firm revenues. These other moments can allow σ and a to be separately identified.

In our discussion of export supply, we noted that the Melitz model implied a straightforward parameterization of the distribution of firm-level sales data. The distribution should also be Pareto distributed, with shape $\zeta = a/(\sigma - 1)$. It is straightforward to estimate ζ on firm-level sales data, but estimation of the parameter on domestic sales data alone does not allow the separate identification of a and σ. Firm-level data that reports sales in multiple markets, however, reveals that there is substantial variation across destinations in the shape of the distribution of origin firm sales. Eaton *et al.* (2011), for example, note that in the French data they employ, high productivity firms enter many more markets than low productivity firms. However, conditional on entry, the distribution of firm sales appears flatter than what one would expect from the productivity differences implied by the entry decision. This empirical analysis suggests that the simple Melitz framework employed by Balistreri *et al.* may need to be complicated if the goal is to replicate observable features of the firm sales distribution.

Arkolakis (2010) provides a rationale for relatively flat sales distributions in destination markets. He introduces increasing marginal costs of market penetration into a model of firm heterogeneity. Firms have available to them an advertising technology that attracts consumers with diminishing probability as the firm's market penetration rises. This introduces an element of diminishing returns, which changes the shape of distribution of firm sales. A reduction in bilateral trade costs will produce relatively larger increases in sales of the less productive active firms. An increase in market size also favors the less productive firms in a relative sense. Arkolakis calibrates the model in order to hit key features of the data.

Eaton *et al.* (2011) derive a very similar model and derive an estimator that can produce estimates of the structural parameters. The French firm-level data they have produces empirical information on n_{ij}, and this together with trade data documenting X_{ij} allows them to observe average firm sales in each market. These are held fixed at their observed levels in estimation. The model contains a fixed trade cost that is dependent on the firm's level of market penetration. This form follows that of the marketing technology in Arkolakis (2010):

$$M(f) = \frac{1 - (1 - f)^{1-1/\lambda}}{1 - 1/\lambda}, \tag{18.32}$$

where f is the fraction of consumers that are reached by each firm and λ is a shape parameter that links market penetration to fixed costs. The Melitz model can be understood in this context as $f = 1$ and $\lambda \to \infty$. Cross country variation in firm-level

penetration allows the authors to pin down an estimate of $\hat{\lambda} = 0.91$, which is notably different than the Melitz benchmark.[49] The procedure also produces an estimate of $\hat{\zeta} = 2.46$.

Recall that ζ represents an amalgamation of the firm heterogeneity parameter a and the elasticity of substitution σ. In order to untangle these parameters, the authors rely on a procedure developed in Bernard *et al.* (2003). That technique introduces other sources of firm size variation, notably idiosyncratic demand shifters in domestic sales. Firm productivity levels predict both larger firm sizes and a greater probability of exporting, whereas demand heterogeneity only affects domestic firm size and not export status. Thus export status serves as an instrument that facilitates the separate identification of a and σ. In an application of this technique to US data Bernard *et al.* (2003) estimate $\hat{a} = 3.6$ and $\hat{\sigma} = 3.8$. In French firm-level data Eaton *et al.* (2011) estimate $\hat{a} = 4.9$ and $\hat{\sigma} = 3.0$.

Cherkashin *et al.* (2010) exploit data from a World Bank survey of Bangladeshi apparel firms to estimate an even richer parametric structure, including destination-specific fixed costs, region-specific demand variation, destination-specific values of σ and the parameters defining a Weibull distribution of firm-level productivity. The richness of the particular data available in this setting allow the authors to estimate a broad number of structural parameters. It is unclear that the parameter estimates extracted in this setting should be taken as informative for the more aggregated, general equilibrium settings that are common in the CGE literature.

18.5.4 Summary

Recently the empirical literature in international trade has adopted estimation approaches that impose general equilibrium conditions as side constraints on the estimation of structural parameters. This is in contrast to linear econometric methods that merely control for general equilibrium effects via the inclusion of fixed effects in estimation. Much of the structural estimation literature has also sought to make quantitative inferences about unmeasured trade costs. In most applications it is difficult to separately identify parameters that define unmeasured trade costs and structural responses to such costs.

One can inform this identification problem in a number of ways. Some researchers impose outside estimates of a structural parameter. Others bring in additional data, such as data on retail prices, and use bilateral variation in such prices to impute bilateral trade costs. Theories of heterogeneous firms introduce yet another parameter into the identification problem. While bilateral trade data is insufficient to allow a complete parameterization of such models, evidence from firm-level trade data has been used to separately identify the key parameters a and σ. Such estimates may serve as a useful benchmark for efforts to calibrate models of firm heterogeneity.

[49] The implication is that small firms face relatively low fixed costs of entering foreign markets, but that the costs of increasing sales in foreign markets are large.

REFERENCES

Anderson, J., van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. Am. Econ. Rev. 93, 170–192.

Alaouze, C.M., Marsden, J.S., Zeitsch, J., 1977. Estimates of the elasticity of substitution between imported and domestically produced commodities at the four digit ASIC level. IMPACT Working Paper 0-11. University of Melbourne, Melbourne.

Arkolakis, C., 2010. Market penetration costs and the new consumers margin in international trade. J. Polit. Econ. 118, 1151–1199.

Arkolakis, C., Costinot, A., Rodríguez-Clare, A., 2010. New trade models, same old gains?. NBER Working Paper 15628 NBER, Cambridge, MA.

Axtell, R.L., 2001. Zipf distribution of US firm sizes. Science 293 (5536), 1818–1820.

Bagwell, K., Staiger, R., 1999. An economic theory of the GATT. Am. Econ. Rev. 89, 215–248.

Baldwin, R., Venables, A., 2010. Relocating the value chain: off-shoring and agglomeration in the global economy. NBER Working Paper 16611. NBER, Cambridge, MA.

Balistreri, E.J., Hillberry, R.H., 2007. Structural estimation and the border puzzle. J. Int. Econ. 72, 451–463.

Balistreri, E.J., Hillberry, R.H., 2008. The gravity model: an illustration of structural estimation as calibration. Econ. Inq. 46, 511–527.

Balistreri, E.J., Hillberry, R.H., Rutherford, T.F., 2011. Structural estimation and solution of international trade models with heterogeneous firms. J. Int. Econ. 83, 95–108.

Bernard, A.B., Eaton, J., Jensen, J.B., Kortum, S., 2003. Plants and productivity in international trade. Am. Econ. Rev. 93, 1268–1290.

Besedes, T., Prusa, T., 2006a. Ins, outs and the duration of trade. Can. J. Econ. 39, 266–295.

Besedes, T., Prusa, T., 2006b. Product differentiation and the duration of US import trade. J. Int. Econ. 70, 339–358.

Broda, C., Greenfield, J., Weinstein, D., 2006. From groundnuts to globalization: a structural estimate of trade and growth. NBER Working Paper 12512. NBER, Cambridge, MA.

Broda, C., Limao, N., Weinstein, D., 2008. Optimal tariffs and market power: the evidence. Am. Econ. Rev. 98, 2032–2065.

Broda, C., Weinstein, D., 2006. Globalization and the gains from variety. Q. J. Econ. 121, 541–585.

Brown, D., 1987. Tariffs, the terms of trade and national product differentiation. J. Policy Model. 9, 503–526.

Chaney, T., 2008. Distorted gravity: the intensive and extensive margins of international trade. Am. Econ. Rev. 98, 1701–1721.

Chen, N., Novy, D., 2011. Gravity, trade integration, and heterogeneity across industries. J. Int. Econ. 85, 206–221.

Cherkashin, I., Demidova, S., Kee, H.L., Krishna, K., 2010. Firm heterogeneity and costly trade: a new estimation strategy and policy experiments. NBER Working Paper 16557. NBER, Cambridge, MA.

Choi, Y.C., Hummels, D., Xiang, C., 2009. Explaining import quality: the role of the income distribution. J. Int. Econ. 78, 293–303.

de Melo, J., Robinson, S., 1989. Product differentiation and the treatment of foreign trade in computable general equilibrium models of small economies. J. Int. Econ. 27, 47–67.

Diewert, W.E., Morrison, C.J., 1986. Export supply and import demand functions. NBER Working Paper 2011. NBER, Cambridge, MA.

Eaton, J., Kortum, S., 2002. Technology, geography, and trade. Econometrica 70, 1741–1779.

Eaton, J., Kortum, S., Kramarz, F., 2011. An anatomy of international trade: evidence from French firms. Econometrica 79, 1453–1498.

Eby-Konan, D. and Maskus, K., 1996. A computable general equilibrium analysis of Egyptian trade liberalization scenarios, University of Hawaii Working Paper 97-1. University of Hawaii.

Erkel-Rousse, H., Mirza, D., 2002. Import price elasticities: reconsidering the evidence. Can. J. Econ. 35, 282–306.

Fair, R.C., 1970. The estimation of simultaneous equation models with lagged endogenous variables and first order serially correlated errors. Econometrica 38, 507–516.

Feenstra, R., 1994. New product varieties and the measurement of international prices. Am. Econ. Rev. 84, 157–177.
Feenstra, R.C., 2010. Measuring the gains from trade under monopolistic competition. Can. J. Econ. 43, 1–28.
Feenstra, R.C., Weinstein, D.E., 2010. Globalization, markups, and the US price level. NBER Working Paper 15749. NBER, Cambridge, MA.
Gallaway, M.P., McDaniel, C.A., Rivera, S.A., 2003. Short-run and long-run industry-level estimates of US Armington elasticities. N. Am. J. Econ. Finance 14, 49–68.
Goldberg, P., Maggi, G., 1999. Protection for sale: an empirical investigation. Am. Econ. Rev. 89, 1135–1155.
Grossman, G., Helpman, E., 1994. Protection for sale. Am. Econ. Rev. 84, 833–850.
Hallak, J.C., 2006. Product quality and the direction of trade. J. Int. Econ. 68, 238–265.
Hallak, J.C., Schott, P., 2011. Estimating cross-country differences in product quality. Q. J. Econ. 126, 417–474.
Handley, K., 2011. Exporting under Trade Policy Uncertainty: Theory and Evidence. Stanford University, Stanford, CA.
Handley, K., Limao, N., 2010. Trade and Investment under Policy Uncertainty: Theory and Firm Evidence. University of Maryland, College Park, MD.
Harrison, G.W., Vinod, H.D., 1992. The sensitivity analysis of general equilibrium models: completely randomized factorial sampling designs. Rev. Econ. Stat. 74, 357–362.
Harrison, G.W., Jones, R., Kimbell, L.J., Wigle, R., 1993. How robust is applied general equilibrium analysis? J. Policy Model. 15, 99–115.
Haynes, S.E., Stone, J.A., 1983. Specification of supply behavior in international trade. Rev. Econ. Stat. 65, 626–631.
Helpman, E., Melitz, M., Rubinstein, Y., 2007. Estimating trade flows: trading partners and trading volumes. Q. J. Econ. 123, 441–487.
Hertel, T., Hummels, D., Ivanic, M., Keeney, R., 2007. How confident can we be of CGE-based assessments of free trade agreements? Econ. Model. 24, 611–635.
Hillberry, R., Anderson, M., Balistreri, E., Fox, A., 2005. Taste parameters as model residuals: assessing the "fit" of an Armington trade model. Rev. Int. Econ. 13, 973–984.
Hillberry, R., Hummels, D., 2008. Trade responses to geographic frictions: a decomposition using micro-data. Eur. Econ. Rev. 52, 527–550.
Hummels, D., 2001. Toward a Geography of Trade Costs. Purdue University, Purdue, IN.
Hummels, D., Klenow, P., 2005. The variety and quality of a nation's exports. Am. Econ. Rev. 95, 704–723.
Hummels, D., Skiba, A., 2004. Shipping the good apples out: an empirical confirmation of the Alchian–Allen conjecture. J. Polit. Econ. 112, 1384–1402.
Hummels, D., Lugovskyy, V., 2009. International pricing in a generalized model of ideal variety. J. Money Credit Bank. 41, 3–33.
Hummels, D., Lugovskyy, V., Skiba, A., 2009. The trade reducing effects of market power in international shipping. J. Dev. Econ. 89, 84–97.
Jacks, D., Meissner, C., Novy, D., 2011. Trade booms, trade busts, and trade costs. J. Int. Econ. 83, 185–201.
Jomini, P., McDougall, R., Watts, G., and Dee, P. S., 1994. The SALTER Model of the World Economy: Model Structure. Database and Parameters, Industry Commission, Canberra.
Kee, H.L., Nicita, A., Olarreaga, M., 2008. Import demand elasticities and trade distortions. Rev. Econ. Stat. 90, 666–682.
Kehoe, T. J., Ruhl, K.J., 2009. How important is the new goods margin in international trade?. Staff Report 324 Federal Reserve Bank of Minneapolis, Minneapolis, MN.
Khandelwal, A., 2010. The long and short of quality ladders. Rev. Econ. Stud. 77, 1450–1476.
Kohli, U., 1978. A gross national product function and the derived demand for imports and supply of exports. Can. J. Econ. 11, 167–182.
Kohli, U., 1993. A symmetric normalized quadratic GNP function and the US demand for imports and supply of exports. Int. Econ. Rev. 34, 243–255.

Krugman, P., 1980. Scale economies, product differentiation and the pattern of trade. Am. Econ. Rev. 70, 950–959.
Krugman, P., 1991. Increasing returns and economic geography. J. Polit. Econ. 99, 483–499.
Krugman, P., Venables, A., 1996. Integration, specialization and adjustment. Eur. Econ. Rev. 40, 959–967.
Lancaster, K., 1979. Variety, Equity, and Efficiency: Product Variety in an Industrial Society. Columbia University Press, New York.
Leamer, E., 1981. Is it a demand curve, or is it a supply curve? partial identification through inequality constraints. Rev. Econ. Stat. 63, 319–327.
McCallum, J., 1995. National borders matter: Canada–US regional trade patterns. Am. Econ. Rev. 85, 615–623.
Melitz, M.J., 2003. The impact of trade on aggregate industry productivity and intra-industry reallocations. Econometrica 71, 1695–1725.
Melitz, M.J., Ottaviano, G.I.P., 2008. Market size, trade, and productivity. Rev. Econ. Stud. 75, 295–316.
Pagan, A.R., Shannon, J.H., 1987. How reliable are ORANI conclusions? Econ. Rec. 63, 33–45.
Powell, A.A., Gruen, F.H.G., 1968. The constant elasticity of transformation and the linear supply system. Int. Econ. Rev. 9, 315–328.
Reinert, K.A., Roland-Holst, D.W., 1992. Armington elasticities for United States manufacturing sectors. J. Policy Model. 14, 631–639.
Romalis, J., 2007. NAFTA's and CUSFTA's impact on international trade. Rev. Econ. Stat. 89, 416–435.
Ruhl, K., 2008. The International Elasticity Puzzle. University of Texas-Austin, Austin, TX.
Rossi-Hanberg, E., 2005. A spatial theory of trade. Am. Econ. Rev. 95, 1464–1491.
Rutherford, T.F., Rutsrom, E.E., Tarr, D., 1993. Morocco's free trade agreement with the European Community: a quantitative assessment. World Bank Policy Research Working Papers 1173. World Bank, Washington, DC.
Schott, P.K., 2004. Across-product versus within-product specialization in international trade. Q. J. Econ. 119, 647–678.
Shiells, C.R., Stern, R.M., Deardorff, A.V., 1986. Estimates of the elasticities of substitution between imports and home goods for the United States. Weltwirtschaftliches Archiv 122, 497–519.
Shoven, J., Whalley, J., 1972. A general equilibrium calculation of the effects of differential taxation of income from capital in the US. J. Publ. Econ. 1, 281–321.
Simonovska, I., Waugh, M., 2011. The elasticity of trade: estimates and evidence. UC Davis Working Paper 11-2. UC Davis, Davis, CA.
Soderbery, A., 2010. Investigating the asymptotic properties of import elasticity estimates. Econ. Let. 109, 57–62.
Su, C.-L. and K. Judd, 2011. Constrained optimization approaches to estimation of structural models, Econometrica forthcoming.
Trefler, D., 1993. Trade liberalization and the theory of endogenous protection: an econometric study of US import policy. J. Polit. Econ. 101, 138–160.
US ITC, 1989. The economic effects of significant US import restraints: case 1: manufacturing. US ITC Publication 2222. US International Trade Commission, Washington, DC.
US ITC, 2002. The economic effects of significant US import restraints: third update. US ITC Publication 3519. US International Trade Commission, Washington, DC.
US ITC, 2004. US–Australia free trade agreement: potential economy-wide and selected sectoral effects. US ITC Publication 3697. US International Trade Commission, Washington, DC.
Valenzuela, E., Anderson, K., Hertel, T., 2008. Impacts of trade reform: Sensitivity of model results to key assumptions. Int. Econ. Econ. Policy 4, 395–420.
Viner, J., 1950. The Customs Union Issue, Carnegie Endowment for International Peace, New York.
Wigle, R., 1991. The Pagan–Shannon approximation: Unconditional systematic sensitivity in minutes. Empir. Econ. 16, 35–49.
Yi, K.-M., 2010. Can multi-stage production explain the home bias in trade? Am. Econ. Rev. 100, 364–393.