Federal Reserve Bank of Minneapolis
Research Department Staff Report 537

April 2017
First version: September 2016

Quantitative Trade Models: Developments and Challenges*

Timothy J. Kehoe

University of Minnesota,
Federal Reserve Bank of Minneapolis,
and National Bureau of Economic Research

Pau S. Pujolàs

McMaster University

Jack Rossbach

Georgetown University Qatar

ABSTRACT

Applied general equilibrium (AGE) models, which feature multiple countries, multiple industries, and input-output linkages across industries, have been the dominant tool for evaluating the impact of trade reforms since the 1980s. We review how these models are used to perform policy analysis and document their shortcomings in predicting the industry-level effects of past trade reforms. We argue that, to improve their performance, AGE models need to incorporate product-level data on bilateral trade relations by industry and better model how trade reforms lower bilateral trade costs. We use the least traded products methodology of Kehoe et al. (2015) to provide guidance on how improvements can be made. We provide further suggestions on how AGE models can incorporate recent advances in quantitative trade theory to improve their predictive ability and better quantify the gains from trade liberalization.

JEL Codes: F11, F13, F14, F17

Keywords: applied general equilibrium; trade liberalization; input-output linkages; Armington elasticities; trade costs.

*This paper has been prepared for publication in the *Annual Review of Economics*. We thank Kim Ruhl for extensive discussions of the issues addressed in this paper. An online data appendix is available at http://www.econ.umn.edu/~tkehoe. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Since the 1980s, applied general equilibrium (AGE) models — sometimes referred to as computable general equilibrium models — have served as the tool of choice for evaluating the economy-wide impact of changes in trade policy. Originally defined as any general equilibrium model that researchers solve numerically after calibrating its parameters to data, AGE models now distinguish themselves from other common quantitative trade models — most of which now fit the prior definition — by their multi-industry and multi-country nature and by their focus on input-output (IO) linkages across industries.

In recent decades, improvements in computation capability and increases in data availability have lowered the cost of building and using these models. This has played a substantial role in cementing the status of AGE models as the standard methodology for evaluating the impact of trade policy. Furthermore, the detailed production structure and linkages featured in AGE models allow them to predict changes in industry-level production and trade flows in response to trade reforms, and these industry-level changes are typically the focus of policy discussion surrounding the desirability of different trade policies.

Although AGE models have remained prominent in policy analysis, their theoretical development has slowed significantly in recent years, as the academic trade literature has shifted its attention to firm-level data and models that focus on them. We intend this paper as a guide to where and why AGE models demand more attention among researchers. Despite their widespread use in policy analysis, AGE models do not have a good track record in predicting the impact of trade reforms on production and trade flows by industry. We hypothesize that the performance of AGE models can be substantially improved by making two major modifications, each of which requires improvements in both theory and measurement. First, we need to improve the theoretical mechanism by which imports and domestic outputs substitute for each other and use micro data to measure the degree of substitutability. In the language of AGE modeling, we need better estimates of Armington elasticities. We use the least traded products (LTP) methodology of Kehoe et al. (2015), which focuses on product-level data on bilateral trade relations by industry, to provide guidance on how improvements can be made. Second, we need to improve our theory and measurement of how trade reforms lower bilateral trade costs.

We have organized the paper around three themes. First, in Sections 2 and 3, we review the development and use of AGE models and outline how researchers can use these models to evaluate the economic impact of trade policy reforms. Second, in Section 4, we review and expand on research that evaluates the performance of AGE models in predicting the effects of past trade reforms, and argue that these models have performed poorly in predicting changes in industry-level production and trade patterns. In particular, we find that industries composed of products that are traded in small yet positive amounts expanded much more following trade liberalization than was predicted by the models. Third, in the remaining sections, we discuss recent developments in the academic trade literature and evaluate the extent to which incorporating these advances into AGE models can overcome the shortcomings identified in our paper.

We intend the message of our paper to be hopeful. Although our results suggest that AGE models have not performed well in identifying which industries will expand the most following trade liberalization, the trade literature has advanced significantly since the development of AGE models in the 1980s. We hypothesize that incorporating recent advances in the literature into AGE models will allow us to develop models with improved accuracy and reliability for trade policy analysis. It is essential that AGE modelers continuously reevaluate the performance of their models using the sort of methodology we employ in this paper. Doing so will allow AGE modelers to identify where their models perform well and where additional research is needed.

2. Development and Use of Applied General Equilibrium Models

AGE models enjoyed a golden age in the academic international trade literature that lasted from the early 1980s until the mid-1990s. As Kehoe & Prescott (1995) explain, "Applied general equilibrium analysis is defined to be the numerical implementation of general equilibrium models calibrated to data: An applied GE model is a computer representation of a national economy or a group of national economies, each of which consists of consumers, producers, and possibly a government." In international trade, AGE models are used extensively to provide quantitative estimates on the economic impact of policy reforms, particularly trade liberalization. Dervis et al. (1982) and Shoven & Whalley (1984) provide surveys of the early AGE literature, whereas Whalley (1985), Shoven & Whalley (1992), and Kehoe & Kehoe (1994a) provide guidance on how AGE models can be used to evaluate the impact of trade liberalizations.

A typical AGE model of international trade consists of multiple countries that trade with each other; each country contains multiple industries, all of which are linked through an IO

structure. In each industry, capital and labor are combined, often using a Cobb-Douglas technology, to produce the industry's value added. This is then combined with intermediate inputs from other industries, often via a fixed coefficients technology, to produce the industry's gross output. A representative household in each country maximizes a utility function, usually homothetic, defined over the consumption of goods from each industry, as well as over public consumption and investment goods. The household purchases goods using the income received from renting its labor and capital to firms and from the resources generated by tariffs on imported goods. Finally, models have market clearing conditions that pin down all wages and prices in the model.

AGE models are designed from the ground up to be able to reproduce key features of the data for the countries of interest. Introducing investment goods into the utility function, for example, makes a static model consistent with the investment that appears in the IO tables. Likewise, models must be consistent with the fact that goods in most industries are both imported and exported simultaneously; introducing individual preferences and production function specifications consistent with the Armington specification (named after Armington (1969)) allows the model to deliver this desired pattern while remaining tractable enough to allow for differing elasticities of substitution across industries.

While AGE models were being developed, the trade literature as a whole was experiencing a renaissance. After Grubel & Lloyd (1971) pioneered the analysis of intra-industry trade, trade theory evolved from featuring perfectly competitive environments to accommodating imperfect competition and scale economies. Krugman (1980) developed the first monopolistically competitive model to rationalize intra-industry trade using Dixit & Stiglitz (1977) preferences. These advances were adapted into AGE models by Harris (1984) in a small open economy model of Canada and were later expanded by Smith & Venables (1988) to study the impact of removing trade barriers in the European Community in a preliminary assessment of what would become the Single Market in Europe. These studies highlight the long-standing tradition of embedding advances from trade theory into AGE models to better understand the impact of policy reform and answer the questions that are important to policy makers.

Many significant advancements in AGE models were developed by researchers analyzing the impact of the North American Free Trade Agreement (NAFTA), the largest free trade area in the world at the time it was enacted in 1994. For instance, the first AGE model to incorporate

Dixit & Stiglitz (1977) preferences was developed by Brown & Stern (1989), who used the model to evaluate the impact of the Canada-U.S. Free Trade Agreement (CUSFTA, precursor of NAFTA) on consumers' welfare, focusing their analysis on the expanded set of varieties available to consumers. That NAFTA led to a significant leap forward for the AGE literature is highlighted by Kehoe & Kehoe (1994b), who compare and summarize the contributions and modeling choices behind several studies, including those of Brown et al. (1992), Cox & Harris (1992), Markusen et al. (1995), and Sobarzo (1995), all of whom use AGE models to study the impact of NAFTA.

Following NAFTA, the number of articles published in peer-reviewed journals by academic researchers using AGE models to assess the impact of trade reforms has diminished. [Recent exceptions include Li & Whalley (2014), who study the impact of the Trans-Pacific Partnership (TPP) on China, and Caliendo & Parro (2015), which we discuss in detail in Section 5.2.] Despite the decline in academic research on AGE models, they have remained heavily used in policy work.[1] For example, Narayanan et al. (2016) review the modeling strategies used by policy makers to evaluate a number of recent trade agreements and show that AGE models remain the dominant methodology that policy makers use to evaluate the impact of a wide range of policy reforms.

The tension between the prevalence of AGE models in policy work and their relative absence in academic research is what drives us to write this paper. One interpretation of the lack of research combined with their widespread use in policy could indicate that AGE models have succeeded and do not require additional development; they work for their intended uses. As we argue starting in Section 4, however, this is not true. Kehoe et al. (2015) show that AGE models fall short when it comes to correctly identifying the industries in which we should expect the positive gains from trade to materialize, one of the primary purposes of using AGE models in the first place. We do not think this indicates that policy makers should move away from AGE models. Indeed, there do not appear to be any realistic alternatives for answering the types of policy questions that AGE models are designed to address. Rather, our claim is that the way forward

[1] One area in the academic literature in which AGE models have thrived is environmental economics. This literature was pioneered by Hazilla & Kopp (1990), who assessed the regulations mandated by the Clean Air Act and the Clean Water Act, and Grossman & Krueger (1994), who assessed the environmental impact of NAFTA by combining an AGE model with industry-level pollution estimates. Since then, a number of AGE models of international trade and the environment have been developed to study the impact of various environmental reforms, especially multilateral environmental agreements involving many countries. Examples include the work of Burniaux & Truong (2002), Peterson et al. (2011), and the survey by Böhringer & Löschel (2006).

should be a return to the tradition of bringing advances from other areas of trade theory into AGE modeling. In the sections that follow, we describe how AGE models are used in practice, how their performance can be evaluated, and how we hypothesize that they can be improved.

3. Using AGE Models to Predict the Industry-Level Impact of Trade Reform

Before discussing the construction and calibration of AGE models, it is important to recognize that researchers have a wide range of modeling choices, as well as different calibration strategies, within the umbrella of AGE models. Given this flexibility, it is not always clear how much these choices matter for the performance of AGE models. Kehoe et al. (1995) evaluate the impact that various modeling choices have on the performance of AGE models in predicting a Spanish tax reform and find that the results are largely unchanged across common alternative specifications. This finding is consistent with the evaluation of several different NAFTA studies by Kehoe (2005), who finds that they perform similarly despite different modeling assumptions and calibration strategies.

This finding is important for our argument because it implies that, when we evaluate the performance of AGE models, we should expect our evaluation to apply to a large class of these models rather than to only the individual models that we consider. This begs the question: How can we hope to improve the performance of AGE models if different implementations perform similarly? The answer is that two elements are essential to the success of AGE models: To use AGE models in a predictive sense, we need to make sure we have the correct specification of how parameters change with trade reforms, that is, the correct specification of the shocks to introduce into the model and the correct elasticities to evaluate a given change in parameters. Because we need to understand the mechanics behind AGE models to estimate shocks and elasticities, we first discuss the structure of AGE models and the calibration of their other parameters.

3.1. Construction and Calibration of AGE Models

The standard procedure for calibrating the parameters of an AGE model (with the exception of policy shocks and elasticities, the calibration of which we discuss in later sections) relies on using equilibrium relationships that hold in the model to pin down parameter values using the data for a given base period. This means that the model exactly matches the data in the base period and that many of the parameters of AGE models can be estimated in a straightforward way from the national accounts and IO data. Total supplies of factors are determined by normalizing the base

year prices of each factor to one and then setting total supplies of labor, capital, and intermediate inputs equal to their factor payments from the national accounts. Utility functions and production functions are frequently assumed to have Cobb-Douglas or fixed-coefficient forms, which makes it easy to calibrate factor and demand intensities directly from IO tables. As a simple example, Kehoe & Kehoe (1994a) show how to calibrate a small AGE model using a three-industry IO table under different specifications of competition and returns to scale.

Traditional sources for IO tables include individual countries' national accounts (the Bureau of Economic Analysis produces a set of IO tables for the United States, for example) and the Organisation for Economic Co-operation and Development (OECD) Structural Analysis (STAN) Input-Output Database for OECD member countries. These sources come with limitations in their coverage of countries, and, when using nationally constructed IO tables, differences in the methodologies used to construct the tables limit the ease with which the national IO tables can be compared and combined for use in calibrating AGE models. More recently, the World Input-Output Database (WIOD), described by Timmer et al. (2015), and the Global Trade Analysis Project (GTAP), described by Hertel (2013), have greatly increased the number of standardized IO tables available to researchers. The WIOD is a publically available dataset that currently covers 40 countries and 35 industries over the period 1995–2011. The GTAP database has the largest coverage of countries, with the GTAP 9 database covering 140 countries and 57 industries for 2004, 2007, and 2011; however, unlike the other sources, the most recent version of the database is not freely available and must be purchased for use. (GTAP databases that are at least two releases old become freely available for download.) We are not aware of any evidence that one of the databases should be preferred to the other. We therefore expect that differences in cost and coverage should dictate which database a researcher should use.

The Armington specification of AGE models, according to which output in each industry is differentiated by country of origin (cars produced in Germany, for example, are different from cars produced in Japan), ensures that production of final outputs uses both domestic and foreign intermediate inputs and that consumers consume both domestic and foreign final output goods. IO data are typically not able to distinguish between countries of origin on an industry-use basis. To give an example, we know how much machinery is imported from Canada in the United States, but we do not know in which industries this machinery is then used. To get around this limitation and reduce the scale of the model, we assume that the aggregation takes place at the border, or

before the inputs are used in the production of other industry outputs. This means that U.S. imports of machinery from Canada are combined with machinery from other countries via an Armington aggregator to get a U.S. machinery aggregate, which is then used in the production of output for other industries in the United States according to the production functions specified above. Although the assumption that industry output is differentiated by country of origin may appear ad hoc, the Armington assumption leads to a parsimonious model that is easy to calibrate and, as we discuss below, shares many properties with a wide class of models.

The Armington aggregator, which combines output from each source country into a single industry aggregate, is a constant elasticity of substitution (CES) function

$$y_{jk} = \theta_{jk} \left(\sum_{i=1}^{m} a_{ijk} y_{ijk}^{\rho_k} \right)^{1/\rho_k} . \tag{1}$$

In this equation, y_{ijk} is the imports by country j of the output of industry k from country i, y_{jk} is the aggregate amount of output of industry k available in country j for consumption or use as an intermediate input in production, and a_{ijk} are non-negative parameters that govern the relative demand for each good. The elasticity of this CES function, $1/(1-\rho_k)$, referred to as the Armington elasticity, determines the degree of differentiation across origins and, as we mention above, plays a crucial role in determining the response of trade flows and consumption patterns to changes in productivity parameters and trade costs. AGE models typically take Armington elasticity estimates from elsewhere in the trade literature because they cannot be calibrated using IO tables and national accounts data but instead require intertemporal data on how trade flows respond to changes in trade costs. Given any elasticity estimate, the share parameters, a_{ijk}, can then be calibrated by forcing the model to match observed foreign and domestic expenditure shares.

Armington elasticities can be calibrated or estimated via several methods, including strategies based on models that are similar to an Armington model yet come with alternative interpretations for the elasticity. As shown by Feenstra et al. (2014), Armington elasticity estimates can be sensitive to the choice of estimation strategy. Despite improvements in understanding the role that the Armington elasticity estimates play in determining the effects of policy reforms in the model, the literature has yet to converge to a single calibration or estimation strategy. In Section 4.2, we evaluate an AGE model that uses standard Armington elasticity

estimates from the literature and show that it performs poorly in predicting the industry-level changes in trade flows following liberalization. We also provide evidence that improved elasticity estimates could improve the performance of AGE models, and, in Section 5, we discuss how recent advances from the theoretical trade literature might be used toward this end.

Because most AGE models are static, aggregate trade imbalances are imposed as an exogenous parameter to match observed aggregate trade imbalances. Whereas overall trade imbalances are exogenously imposed, the industry and bilateral compositions of these trade imbalances arise endogenously and will change in response to policy reform. That is, an AGE model of the United States would impose an overall trade deficit, but it would not impose the individual countries that the United States has bilateral deficits with, nor would it impose the industries in which net exports are negative. Although it is difficult to endogenize aggregate trade imbalances in static models (or current account imbalances in models that feature international lending), Dekle et al. (2008) and Kehoe et al. (2013) show that removing aggregate current account imbalances is likely to lead to large redistributional effects across traded and nontraded industries despite a muted overall impact on real GDP.

A robust empirical observation is that countries tend to disproportionately consume domestic output compared with consumption of foreign varieties of the same goods, as well as disproportionately using domestic intermediate inputs in production. Although this home bias is commonly thought to arise because of trade costs and barriers that limit international trade (see Obstfeld & Rogoff (2001)), these elements are often not modeled explicitly in AGE models. Standard AGE models instead rationalize observed home bias through the share parameters a_{ijk} of the Armington aggregators in the model, effectively assuming home bias exists solely because of preferences and production technologies. A shortcoming of this strategy is that it fails to leave open the possibility that the degree of home bias may change in response to changes in trade policy. For that reason, recent studies have focused on better understanding the possible origins of home bias. Two recent and influential studies discussing the home bias puzzle are those of Hanson & Xiang (2004), who present a model of endogenous product differentiation where home bias is generated by variation in transportation costs across industries, and Yi (2010), who investigates the role of vertical linkages and fragmentation of production in generating home bias. Wolf (2000) additionally points out that home bias (or local bias) also exists on a subnational level.

As home bias has become better understood, researchers have continued to shift away from using the Armington specification to capture home bias, instead opting to explicitly model the trade costs that lead to a home bias pattern. Corsetti et al. (2007) provide an example in which incorporating home bias through trade costs is quantitatively important for the transmission and welfare impact of macroeconomic shocks. It is worth noting, however, that current methods for estimating model-derived trade costs do not completely resolve the shortcomings of embedding home bias in preferences and production functions. The reason for this is that these methods are typically not able to identify the exact components of these trade costs or the degree to which these trade costs will change in response to changes in trade policy. Therefore, although model-derived trade costs are a step in the right direction, more work must be done to make estimates of these costs useful in AGE modeling.

The final set of parameters that must be calibrated is the policy parameters. These parameters are the most important ones in the model: Dhanging these parameters is what allows the model to perform counterfactuals. Although, in principle, any of the exogenous parameters can be changed, in the context of international trade, the most common choices of policy parameters are taxes, trade costs, and industry productivity parameters. Kehoe et al. (2013) evaluate a policy reform that falls outside of those choices, in which the U.S. trade deficit is the policy parameter itself. In the AGE models that are used to evaluate the potential impacts of trade reform, the central policy parameters are typically tariffs and trade costs, the calibration of which we devote the next subsection to. Common trade reforms studied in AGE frameworks include the uniform lowering of all tariffs by a set amount; the complete removal of all tariffs; and, when available, changing tariffs from calibrated initial pre-reform levels to proposed post-reform levels.

Although changes in trade policy are taken as exogenous in AGE models, there is a large body of literature studying the endogeneity of tariffs and trade policy, as well as the political economy of international organizations such as the World Trade Organization (WTO). Bagwell & Staiger (1999) show how two key elements of the WTO, reciprocity and nondiscrimination, can act to offset the terms-of-trade externality and resulting prisoner's dilemma that causes high tariff rates to arise endogenously when nations set tariffs unilaterally. There is also a large body of theoretical literature examining how tariffs and subsidies interact with the industrial structure of countries. For example, Broda et al. (2008) discuss how market power and concentration can affect optimal tariffs for a country, and Costinot et al. (2015) develop a framework to study how

comparative advantage might influence optimal tariffs. In the other direction, Demidova & Rodriguez-Clare (2009) show that export subsidies generate productivity increases. These and related studies suggest that observed tariffs contain information beyond just the tariff rates themselves. Baier & Bergstrand (2007) show that accounting for the endogeneity of trade policy quintuples the estimated impact of free trade agreements on trade flows in a gravity framework. AGE models, however, rarely incorporate this type of information, which is a shortcoming that future research on AGE models should seek to address.

3.2. **Tariffs and Trade Costs for Evaluating Trade Reforms**

Tariff parameters in AGE models can be calibrated in two standard ways. The first is to use observed tariff revenues, as Kehoe & Kehoe (1994a) do, to back out implied tariff rates. This has the benefit of generating tariff revenues that are consistent with the national accounts. Despite a large literature predicting that countries should impose tariffs uniformly across products (see Opp (2010) for a recent example and Costinot et al. (2015) for a notable exception), in practice, tariff rates are rarely set uniformly. To account for this empirical regularity, we can alternatively calibrate product- or industry-level tariff rates in the model to match tariff rate measures provided by the United Nations Conference on Trade and Development (UNCTAD) Trade Analysis Information System (TRAINS) database. The TRAINS database contains reported bilateral tariff rate information for products categorized at the six-digit Harmonized System (HS) level and constructs estimated tariff rates for other classification systems and levels of aggregation.

Tariff agreements organized through the WTO, as well as other bilateral and multilateral trade agreements, typically take the form of caps set on the maximum level of tariff rates that can be applied. In practice, these caps are often nonbinding. Maggi & Rodríguez-Clare (2007) rationalize that trade agreements set caps instead of tariff rates themselves because this allows for large immediate cuts in tariff rates followed by further gradual reductions. When calibrating tariff rates in an AGE model, we therefore use effectively applied tariff rates, which are estimates of the tariff rates actually applied (typically the lowest of several caps governed by different trade agreements). Because effectively applied tariffs are reported as ad valorem equivalents, these rates can be inserted directly into the AGE model in the form of tax wedges for consumers and producers.

Most of the countries in the world are now members of the WTO, and significant gains have been made in lowering tariff rates over the past 30 years. Because of this decline in tariffs,

many of the quantitative trade models in the literature, with the exception of AGE models, have moved away from using tariffs as the object of policy reform and focused more generally on trade costs. Iceberg trade costs, as modeled by Samuelson (1954), enter into the AGE framework almost identically to tariffs, with the exception that tariffs produce revenue that is rebated to consumers or used to fund government expenditures, whereas iceberg trade costs are lost completely. This means that there are larger gains from reducing trade costs than there are from reducing tariffs, an implication that is explored quantitatively by Felbermayr et al. (2015).

Relative to tariffs, nontariff trade costs are a less well-defined concept and more difficult to observe directly in the data. One exception is data on freight costs charged by shipping providers to transport goods internationally. Recent papers have begun making use of this data (see, for example, the studies by Adao et al. (2015) and Shapiro (forthcoming)). Other papers, however (e.g., by McCallum (1995) and Wolf (2000)), show that freight costs do not appear to make up a large fraction of observed trade costs. Recent research has also highlighted the fact that freight costs are endogenous and should change in response to changes in tariffs. Further, this line of research, for example by Kleinert & Spies (2011) and Asturias (2016), has shown that there is little correlation between traditional trade cost measures, which use distance as a proxy, and freight costs themselves.

Additional components of trade costs include nontariff trade barriers and regulatory policies (Dean et al. (2009), Goldberg & Pavcnik (2016)), marketing costs involved in serving additional markets (Arkolakis (2010)), transportation costs (Limão & Venables (2001)), transportation time (Hummels & Schaur (2013)), exchange rate hedging costs (Allayannis & Ofek (2001)), and fixed costs associated with gaining access to foreign markets (Krugman (1980), Melitz (2003)). Other costs are less obviously trade costs but can nonetheless function as such in trade models; recently proposed examples are information frictions (Allen (2014)) and credit constraints (Manova (2013), Leibovici (2015), Kohn et al. (2016)). The trade costs mentioned here are far from exhaustive — Anderson & van Wincoop (2004) provide a survey of the literature surrounding trade costs — yet they reinforce our point that constructing nontariff trade costs directly from data is a challenging endeavor, prohibitively so for many early AGE applications.

For this reason, rather than directly calibrating trade cost parameters from the data, researchers often calibrate them indirectly using the same model-implied equilibrium relationships used to estimate production function intensities. Anderson & van Wincoop (2003) show that, with

Armington models, implied ad valorem trade costs can be recovered from a gravity regression. Structurally connecting the results from the gravity regression to the Armington framework, they estimate the implied trade costs generated by the border between the United States and Canada and show that the naïve gravity regression estimates of McCallum (1995) are biased and significantly overstate the trade costs implied by the border. Using model-implied trade costs rather than direct data on observed trade costs has allowed researchers to better understand the effects of economic integration, as surveyed by Donaldson (2015). As recent examples of this approach, Comerford & Rodriguez-Mora (2015) apply this strategy to study the potential costs of secession and independence for Scotland and Catalonia, and Ottaviano et al. (2014) evaluate the potential impact on the United Kingdom of leaving the European Union. In these types of studies, the predicted changes in trade costs resulting from changes in economic integration are typically estimated using coefficients from a gravity regression with an indicator variable for economic integration. This method of estimation, however, fails to take into account the endogeneity of economic integration, and, because there are few examples of such disintegration happening in developed economies, it is difficult to validate whether the predicted changes in trade costs are reasonable estimates of the changes that would occur given dissolution.

Over the past decade, significant improvements have been made in the ability of researchers to calibrate model-implied trade costs and elasticities through gravity regressions. Projects such as the Centre d'Études Prospectives et d'Informations Internationales (CEPII)'s Gravity Database, described by Head et al. (2010), have increased the availability of much of the data necessary for running these gravity regressions. Simultaneously, advances have been made in understanding how to estimate the model-implied equilibrium equations. Egger (2000) shows that these gravity relationships are best estimated using panel data, and Santos Silva & Tenreyro (2006) provide an alternative estimation procedure that estimates the relationship without taking logs and avoids the bias arising from Jensen's inequality. Head & Mayer (2014) provide a survey of the advances related to estimating gravity regressions. Although the typical approach is to derive implied trade costs using aggregate trade flows, Irarrazabal et al. (2015) provide micro-level estimates of trade costs using firm-level trade data while exploring the quantitative implications of modeling additive trade costs as multiplicative trade costs. Additionally, although bilateral symmetry in trade costs is often imposed, Hummels et al. (2009) and Waugh (2010) argue

that trade costs are in fact asymmetric and vary systematically depending on the income of the trading partners.

Economic historians have shown that traditional measures of trade costs have declined significantly over time. Model-implied trade costs, however, have remained substantial even as tariffs have decreased to near zero for many countries and products. Jacks et al. (2008) provide estimates of bilateral trade costs between 1870 and 2000, and Jacks et al. (2011) show that changes in trade costs over this period correlate strongly with aggregate changes in trade flows. Estevadeordal et al. (2003) investigate the extent to which transportation costs, tariffs, and the gold standard explain the rise and subsequent decline in world trade from 1870 to 1939. More recently, Donaldson (forthcoming) studies the reduction of transportation costs associated with railroad construction in colonial India using inter-district price differences as a proxy for trade costs.

Although advances in recovering and calibrating trade costs have expanded the usefulness of AGE models in an evaluative sense, challenges remain in incorporating these advances into AGE models used to predict the impact of trade reforms. Following a policy implementation, post-reform trade costs can be recovered using the gravity approach and inserted into the model to evaluate how well the framework can capture observed changes. It may not be possible, however, to predict the resulting changes in trade costs prior to the policy reform. Overall, although some changes in trade costs, such as changes in tariff schedules, are easily predictable and other changes in trade costs can be inferred by predictable changes in gravity regression variables, much work remains to be done to adapt the calibration methods described above for use in AGE models to yield the predictions that policy questions require.

4. Evaluating the Performance of AGE Models

For AGE models to be useful as predictive tools for policy analysis, it is essential that the models be able to forecast the effects of policy reforms with some degree of accuracy. To evaluate the accuracy of AGE models, we can look at the pre-reform predictions that researchers have produced using AGE models, or the predictions that AGE models would have yielded had they been used pre-reform, and compare them with what actually happened post-reform.

When evaluating the accuracy of predictions of the impact of trade reforms, a key question is how we should account for the many unrelated (to the reform) changes that occur post-reform that influence the data and are not considered in the model predictions. One possibility is to decompose observed changes in the data by source and then compare changes in the data attributed

to the trade reform with the predicted impact of the trade reform. The shortcoming of this strategy is that it is not always clear how to decompose the data, particularly if the external shocks to be accounted for are not clearly identified. If the external shocks are well known, another possibility is to modify the predictions of the model by feeding these shocks into the model along with the original counterfactual. This is the strategy that Kehoe et al. (1995) follow when evaluating the performance of the AGE model used by Kehoe et al. (1988) to predict the impact of a tax reform in Spain. They show that accounting for two external shocks, a worldwide drop in the price of petroleum and a negative shock to domestic agricultural productivity due to a draught, improves the performance of the model predictions in matching observed changes in industry-level output.

An alternative strategy when evaluating model performance is to leave the data and model predictions as they are, but to adjust the baseline for what is considered to be success or failure. A natural way to adjust the baseline is by providing benchmark predictions that can be contrasted with those from the model; then success or failure can be determined by running a horserace between the model under evaluation and the benchmark predictions. This is the approach followed by Kehoe et al. (2015) and is the strategy that we follow in this paper. Although there are potential complications in evaluating models this way, particularly if the changes that occur post-reform are large relative to the reform considered in the model, we choose this method of evaluation because it allows us to show clearly that it is possible to build models that are better able to predict the industry-level impact of trade reform. Furthermore, the benchmark predictions help us to understand where AGE models fall short and how they can be improved.

4.1. Previous Evaluations of AGE Models

A natural test case for evaluating the AGE framework is NAFTA. NAFTA was a significant policy reform between countries that continue to trade heavily with each other, and it attracted a large amount of attention from economists both pre- and post-reform. The effectively applied tariff rates between NAFTA countries fell from an average of almost 9 percent in 1989, before the implementation of both NAFTA and CUSFTA, to near zero shortly after the implementation of NAFTA. At the same time, each of the NAFTA countries has remained as one of the top three trading partners of the other NAFTA countries. NAFTA is also an ideal test case for the accuracy of AGE models because AGE models were the primary models used to inform policy makers on how NAFTA would affect the United States, Canada, and Mexico.

Fox (1999), Kehoe (2005), Shikher (2012), and Kehoe et al. (2015) evaluate how the AGE models originally used to predict the effects of NAFTA did in matching actual changes in bilateral industry-level trade flows, and show that the models did poorly. The models did so poorly, in fact, that the predictions of the models were often negatively correlated with the actual changes observed post-NAFTA. This finding is concerning, because if AGE models cannot get one of the largest and most significant trade reforms in recent history correct, then there may be reasons to doubt the reliability of the AGE models currently being used for evaluating trade policy. An important caveat is that these evaluations are concerned only with the accuracy of AGE models in predicting changes in industry-level trade flows following the implementation of NAFTA and do not discuss the arguments for and against NAFTA, which are reviewed and evaluated by Burfisher et al. (2001).

Why did AGE models do so poorly in predicting the impact of NAFTA? Although we do not know for certain, we investigate this question by examining how AGE models perform in predicting the impact of other trade liberalizations. Further, we search for shortcomings in the use of AGE models by examining features of the data that have yet to be fully incorporated into the models. It is worth noting that the failure of AGE models is potentially limited to their application in evaluating the impact of trade reforms. AGE models have been found to perform well for certain other policy reforms; for example, Kehoe et al. (1995) find that the AGE model used by Kehoe et al. (1988) performed well at predicting the impact on industry-level output and prices of a tax reform in Spain. We argue that one reason for the models' poor performance is due to failing to account for product-level trade within industries, in particular, measures of extensive margin growth.

Several papers suggest that information on disaggregated trade flows may be necessary to match aggregate trade flows. Using an AGE framework as an evaluative tool, Romalis (2007) shows that a significant amount of growth in trade between NAFTA countries occurred by displacing trade from non-NAFTA countries. Romalis also emphasizes that, to accurately predict the impact of trade reforms, it is essential to correctly identify the shocks associated with trade reforms and to correctly estimate the trade elasticities and, moreover, that the use of disaggregated trade data can help along both margins. Trefler (2004) shows that plant-level productivity rose significantly in the industries experiencing the largest tariff cuts following the CUSFTA signed in 1988, whereas low productivity plants reduced employment following the trade liberalization.

Hillberry & McDaniel (2002) show that a significant portion of growth in trade between the United States and Mexico occurred in products that were previously not traded. Kehoe et al. (2015) show that using the extensive margin could have led to better predictions than those of the models originally used, a finding we explore further in this paper. One reason that extensive margin growth was not initially built into AGE models is because of a historical lack of theoretical underpinnings. As we discuss in subsequent sections, the significant theoretical advances over the past 15 years have allowed researchers to better understand how the extensive margin is affected by trade policy. Much progress has also been made in understanding how changes in the extensive margin can be mapped into elasticities that allow AGE models to capture these effects.

Before jumping to the conclusion that current AGE models perform badly because AGE failed to predict the effects of NAFTA, it is fair to ask whether this inaccuracy was a shortcoming only with the particular models used at the time and whether the AGE models currently used in policy evaluation have overcome this shortcoming. AGE models may perform better now, for example, because of improvements in the econometric foundations underlying how elasticities for AGE models are estimated. McKitrick (1998) raises econometric concerns over how functional form assumptions were imposed by the early applications of AGE models, whereas Hertel et al. (2007) discuss improvements in the AGE modeling framework that partially or fully alleviate several related econometric concerns.

4.2. Evaluation of the GTAP Model

To evaluate whether AGE models currently used for policy analysis have improved, we use the GTAP model and database described by Hertel (2013) to evaluate several recent bilateral trade agreements and compare the predictions from the GTAP AGE model with observed changes. We choose the GTAP framework because it is widely used for policy work and targeted toward researchers in policy-oriented institutions such as the WTO and World Bank.

We evaluate the GTAP framework for the following bilateral trade agreements (the year in which the trade agreement began implementation is in parentheses): United States–Australia (2005), United States–Chile (2004), China–Chile (2006), and China–New Zealand (2008). It is worth mentioning that the GTAP framework and its extensions were, in fact, used by policy makers to evaluate the potential impact of many of these free trade agreements prior to implementation [see, e.g., reports by the US International Trade Commission (2003) and New Zealand Ministry of Foreign Affairs and Trade and China Ministry of Commerce (2004)] and they continue to be used

by both member and nonmember nations to evaluate bilateral and multilateral trade agreements such as the TPP (see, for example, Burfisher et al. (2014)). To allow for a consistent evaluation of the GTAP framework for all of the above trade agreements, we generate predictions for the impact of the policy reform in a standardized way for each reform using the GTAP 9 database and standard GTAP model described by Hertel (2013).

The standard version of the GTAP model is an AGE model with perfect competition and constant returns to scale and makes use of multi-industry, multi-country, production, tariff, and IO data to generate predictions for shocks to exogenous parameters. The GTAP database, which we use to calibrate the model, is designed to allow for easy aggregation of industries and countries. For each trade reform, we aggregate the base data into three countries: the two partner countries involved in the bilateral trade agreement and a rest-of-world (ROW) aggregate. The industry-level Armington elasticities in the GTAP model are taken from the literature, specifically Hertel et al. (2007), and range from 1.8 for Minerals Not Elsewhere Classified to 34.4 for Gas, with an average elasticity of approximately 7. To evaluate the accuracy of the GTAP models, we generate counterfactual predictions from the model and compare them with actual changes observed in the data.

We generate counterfactual predictions for each trade reform using 2004 as our reference year for calibration, and we consider the impact of setting tariffs to zero for all commodities traded between the two partner countries while leaving the tariff rates between each country and the ROW aggregate unchanged. A minor shortcoming of setting all tariffs to zero is that, in practice, tariffs and trade barriers are often not fully eliminated for all products following liberalization. Despite these few exceptions, however, the vast majority of products do experience complete or significant reduction. Table 1 reports the simple average (in percent) across six-digit Harmonized System products of effectively applied tariff rates between each country pair in 2002 and 2015 (collected from TRAINS) under the columns titled "Average 2002 tariffs" and "Average 2015 tariffs." As we can see, tariffs significantly declined for each of the trade liberalizations, with post-reform rates near or at zero for all of the countries. A more significant concern is that many of these countries signed multiple bilateral trade agreements over the same period, and so holding tariffs fixed with the ROW is an unrealistic assumption and may be the source of inaccuracies. As we discuss in Section 5.2, however, including all tariff changes leads to only minor improvements in accuracy for the AGE models that predict the impact of NAFTA.

To evaluate the accuracy of the GTAP models, we compare the counterfactual predictions of the model with actual growth observed in the data. The GTAP model generates predictions for the percentage change in the value of exports for each industry following the removal of tariffs. To compute the actual changes in growth for each industry, we use trade data collected from the United Nations Commodity Trade Statistics Database (UN Comtrade) at the six-digit Harmonized System (HS2002) level and aggregate this product-level trade into trade flows for GTAP industries using a concordance provided by GTAP. Trade data are available for 42 of the 57 GTAP industries, although not all industries report positive trade for all importer-exporter pairs and all years. Using the trade data, we compute the percentage changes, $z_{ijk}^{t,t'}$, for exports from country i to country j in industry k between periods t and t' as

$$z_{ijk}^{t,t'} = 100 \times \left(\frac{x_{ijk}^{t'}}{x_{ijk}^{t}} - 1 \right), \tag{2}$$

where x_{ijk}^{t} is exports from country i to country j in industry k in period t, reported in current price USD. Although it is common to also deflate trade flows, for example, by the GDP of the exporter, deflating makes no difference for our results, because we focus on correlations to evaluate how the model does in predicting which industries experience the most growth following the trade reforms. Following Kehoe (2005), we set the base period as two years before our base period for tariffs, because significant changes in trade flows often take place prior to the actual implementation of trade agreements due to announcement effects.

We evaluate the GTAP framework using the weighted correlation between the predicted changes from the model and the observed changes in the data, where we weight using each industry's exports, averaged across the base and final period. The weighted correlation tells us to what extent the GTAP model is able to accurately predict the industries that experience the most growth following liberalization. We use weighting to take into account the fact that policy discussion is typically focused around the impact of liberalization on large industries, and so our evaluation should put more weight on whether the model's predictions are accurate for these large industries. The results we show, however, are qualitatively robust to several other weighting schemes based on pre- and post-reform trade values as well as to using the unweighted correlation that factors deviations the same even for industries that account for very little trade.

Computing the weighted correlation requires the weighted average growth rates for both the actual changes in trade flows, $z_{ijk}^{t,t'}$, and the predicted changes in trade flows from the GTAP model counterfactuals, $\hat{z}_{ijk}^{t,t'}$, as well as their weighted variances and the weighted covariance between them. The formula for the weighted average growth rate for actual changes in trade flows is given by

$$\text{mean}(z)_{ijk}^{t,t'} = \sum_{k=1}^{57} \omega_{ijk}^{t,t'} z_{ijk}^{t,t'} \tag{3}$$

and is used in the calculation of the weighted covariance between actual and predicted changes:

$$\text{cov}(z,\hat{z})_{ijk}^{t,t'} = \sum_{k=1}^{57} \omega_{ijk}^{t,t'} \left(z_{ikj}^{t,t'} - \text{mean}(z)_{ijk}^{t,t'} \right) \left(\hat{z}_{ikj}^{t,t'} - \text{mean}(z)_{ijk}^{t,t'} \right), \tag{4}$$

where the weight used for each industry is the industry's share of exports averaged across the base period ($t = 2004$) and the end period ($t' = 2015$). Specifically, the weights are

$$\omega_{ijk}^{t,t'} = \frac{\left(x_{ijk}^{t} + x_{ijk}^{t'} \right)}{\sum_{k'=1}^{57} \left(x_{ijk'}^{t} + x_{ijk'}^{t'} \right)}. \tag{5}$$

The weighted correlation, which serves as our metric for evaluating model performance, is then given by

$$\rho(z,\hat{z})_{ijk}^{t,t'} = \frac{\text{cov}(z,\hat{z})_{ijk}^{t,t'}}{\sqrt{\text{cov}(z,z)_{ijk}^{t,t'}} \sqrt{\text{cov}(\hat{z},\hat{z})_{ijk}^{t,t'}}}. \tag{6}$$

These weights imply that, when actual growth deviates from predicted growth, the deviation factors more heavily into the weighted correlation for industries that are traded more heavily.

Table 1 reports the weighted correlation between the predicted and actual changes in trade flows between 2002 and 2015 using (6) in the column titled "Correlation of GTAP with data" (the last column, "Correlation of LTP with data," is discussed below). When computing these weighted correlations, for each importer-exporter pair, industries that report zero trade in either 2002 or 2015 are dropped from the sample (replaced with zeros in (5)). We also exclude, as outliers, U.S. exports to Chile in the petroleum industry (due to fracking), and Australian exports of cattle meat to the

United States (due to mad cow disease). We explore the sensitivity of these results to the inclusion of these outliers in the appendix.

Table 1: Comparisons of GTAP and LTP predictions for recent trade liberalizations with data

Exporter[a]	Importer	Average 2002 tariffs	Average 2015 tariffs	Correlation of GTAP with data	Correlation of LTP with data
United States	Australia	4.47	0.00	0.27	0.55
Australia	United States	3.86	0.72	–0.14	0.53
United States	Chile	6.98	0.00	0.08	0.55
Chile	United States	2.83	0.07	0.03	0.48
China	Chile	7.00	0.13	0.14	0.61
Chile	China	11.68	0.49	0.04	0.07
China	New Zealand	4.06	0.04	–0.36	0.61
New Zealand	China	11.72	0.45	–0.09	0.48
Simple average		5.40	0.24	–0.00	0.49

[a]Outliers in data excluded: U.S. exports of petroleum to Chile, Australian exports of beef to the United States.

As we can see from Table 1, the weighted correlations between the predicted changes from the GTAP model counterfactuals and the actual changes in growth are low. A fair question to ask when evaluating the GTAP framework is whether any alternative models could have been expected to perform better. As evidence that we should be able to design AGE models that perform better, we adapt the methodology of Kehoe et al. (2015) to show that including information along only a single dimension — the share of exports of LTP in total exports of each industry — can lead to predictions that outperform the AGE models that take into account information on taxes, production, and IO linkages, yet lack this essential margin.

The motivation behind the use of the share of least traded products to predict the growth of each industry is the observation of Kehoe & Ruhl (2013) that the products that experience the most growth following trade reform and growth episodes are the products that were previously traded, but in small amounts. Kehoe et al. (2015) show that the faster growth of LTP applies broadly across industries and cannot be explained by greater changes in trade costs. This could, for instance, indicate that the elasticity of trade flows with respect to changes in trade costs is

greater for LTP compared to non–LTP, a hypothesis motivated by models featuring extensive margin growth.

To construct the share of LTP in each industry, s_{ijk}^{t}, we use the same six-digit HS2002 data and concordance that we use to construct the actual changes in industry trade for each trade reform. We define the set of LTP by sorting all the products by their average trade value over 2002−2004, starting with the products with the smallest average values of trade. For each product, we compute the cumulative value of trade in 2002 of all products with less trade over the 2002−2004 period, and we classify as the set of LTP the set of goods that accounts for exactly 10 percent of trade in 2002. Sorting over multiple years ensures that we are conservative when we compute the share of growth due to LTP because goods that experience lumpy trade will not be classified as LTP. Products that are not traded in 2002 do not enter into the share of LTP, and so the margin we are focusing on is indeed the set of products that are traded in positive, but very small, amounts. As Kehoe et al. (2015) show, this methodology is robust to using alternative cutoffs — for example, 5 percent or 20 percent of trade instead of 10 percent — as long as the cutoff is not so small that it omits much of the margin we want to focus on or so large that it completely dilutes it. Note that, although the overall share of least traded products is 10 percent of total trade flows by construction, in any particular industry this share may be more or less than 10 percent because the set of LTP is computed by ignoring industries.

The hypothesis for our exercise is that the industries that should experience the most growth are those that are disproportionately composed of LTP. Kehoe et al. (2015) show how the share of LTP in each industry can be combined with the change in tariffs and the trade elasticity from a simple gravity regression to construct level predictions for industry-level changes in trade flows. The predictions constructed using this methodology are linear functions of the share of LTP in each industry, and are therefore perfectly correlated with the share of LTP in each industry for each importer-exporter pair. Because we are only interested in the weighted correlations between predicted and actual changes in industry-level trade flows, we do not need to actually generate predictions to evaluate how they would perform; the share of LTP in each industry is sufficient. In the online data appendix, we evaluate the AGE models according to an additional criterion: the mean absolute percentage error, which requires us to generate the level predictions. We do this following the work of Kehoe et al. (2015) and we show that our results are robust to this alternative metric.

The weighted correlations between the share of LTP and actual changes in industry level-trade are reported under the column titled "Correlation of LTP with data" in Table 1. These weighted correlations are produced by substituting the share of least traded products, s_{ijk}^{t}, for $\hat{z}_{ijk}^{t,t'}$ in equations (4)–(6). The LTP predictions outperform the GTAP predictions in terms of matching actual changes in post-reform trade flows for each of the eight importer-exporter pairs. The GTAP predictions perform the best for U.S. exports to Chile following the U.S.-Chilean free trade agreement; however, even in this case, they perform worse than the average weighted correlations between actual changes and the LTP predictions of 0.49, which is much higher than the average weighted correlation of –0.00 for the GTAP predictions.

The superior performance of the LTP predictions can guide our thinking about what changes might improve the performance of AGE models. In general, AGE models may fall short for three broad reasons. The first is that the structure of the model could be incorrect; in other words, that the Armington specification does not emphasize the relevant dimensions of substitutability across products or that the IO structure does not accurately account for inter-industry linkages. We hypothesize that the shortcomings of AGE models are not inherent and that their structure is not the primary issue in large part because similar models have been found to work well in analyzing reforms in other areas of economics such as public finance. The second potential problem is that the shocks inserted into the models to perform counterfactuals do not correctly identify the changes in trade costs associated with liberalization. We think that this is likely a significant issue, because observable trade costs, such as tariffs and freight costs, have little relationship to the main components of model-implied trade costs. Unfortunately, although work must be done to better understand how changes in model-implied trade costs can be predicted ex-ante, our benchmark predictions do not provide insight into how to resolve this issue. The third potential issue, which our benchmark predictions do address, is that the elasticities could be estimated incorrectly. Indeed, our results suggest that industries with more LTP are more elastic — for instance, because of lower marginal marketing costs for firms with smaller market shares, as theorized by Arkolakis (2010). If this is the case, then industry-level elasticities should differ across bilateral country pairs, which is not the case in any existing applications of AGE models.

Overall, our results in this section show that the AGE models still commonly used to predict trade reform likely suffer from the same accuracy problems that plagued the models that originally performed so poorly for NAFTA. Despite this apparent lack of progress, we hypothesize

that several advances in trade theory over the past 15 years have not yet been fully integrated into multi-industry AGE models, and these advances have the potential to improve the reliability of AGE models for policy evaluation. In the sections that follow, we discuss advances in the theoretical trade literature that are relevant to the shortcomings we identify for AGE models. We evaluate a state-of-the-art AGE model that embeds some of these advances into its structure and calibration, and we summarize our view of where the literature should focus in the future and the reasons for our optimism.

5. The Extensive Margin Revolution in International Trade

Although it is not clear why the development of AGE models became less prominent within the academic literature — we hypothesize that it was due to complacency based on the view that these models perform well — it is clear that the emergence of detailed datasets on plants and firms in a number of countries has played a pivotal role in the refocusing of the international trade literature on firm heterogeneity. Early studies examined firms in Chile (Tybout et al. (1991); Pavcnik (2002)), Colombia (Roberts & Tybout (1997)), and South Korea (Feenstra et al. (1999)) and were useful for uncovering many important insights about the role of the firm in international trade. These and related studies investigated, for instance, how firms make decisions about whether to export and how firms respond to trade reforms. Furthermore, many patterns were discovered that helped to guide the theoretical trade literature. For example, even for heavily export-oriented economies, only a small fraction of firms are exporters. Moreover, firms that export tend to be larger, tend to employ more educated workers, and tend to be more productive than firms that do not export.

One of the findings that most directly influenced the development of the theoretical trade literature is that of extensive margin growth and redistribution of inputs among firms. Specifically, after a country undergoes a trade liberalization, some of the firms that previously served only the domestic market begin exporting while others cease production entirely. Hummels & Klenow (2005) measure the extensive margin for 126 exporting countries and find that it accounts for approximately 60 percent of the greater exports of larger economies. Hillberry & McDaniel (2002) find that most of the increase in trade in the United States after NAFTA consists of new varieties coming from Mexico. Kehoe & Ruhl (2013) consider a product-level decomposition of the extensive margin and show that large changes in trade are disproportionately driven by growth in

products that are initially traded in small amounts, which is the inspiration for our benchmark predictions when evaluating the AGE models.

5.1. New Trade Models, New Trade Elasticities

Inspired by these empirical studies of how firms react to trade liberalization, the trade literature developed theoretical models to help us understand the origins and implications of these newly discovered patterns. The most influential of these new models were the Melitz (2003) model of firm heterogeneity with monopolistic competition and trade, and the multidimensional Eaton & Kortum (2002) model of perfect competition and international trade. Arkolakis et al. (2012) show that these new models — as part of the wider class of gravity models, so named because they generate trade flows that are consistent with the gravity regressions discussed in Section 3.2 — are equivalent to the older Armington models in the sense that their welfare predictions depend only on the trade elasticity and the change in domestic trade shares. Simonovska & Waugh (2014) show, however, that despite identical predictions for changes in trade flows for a given fixed trade elasticity and shock to trade costs, the models have different implications for how trade elasticities should be estimated. Before discussing these new alternative methods, we review how Armington elasticities were historically estimated for AGE models.

The earliest applications of AGE models often relied on reasonable guesses for trade elasticities, such as those compiled by Stern et al. (1976). The first attempts to systematically estimate elasticities relied on estimating partial equilibrium demand systems relating changes in prices to changes in trade flows — for instance, the stock-adjustment model employed by Shiells et al. (1986) — and generally delivered estimates in line with the earlier guesses. These early estimates, however, were often considered to be too low, as they generated what some viewed as unrealistically large terms of trade effects. As discussed by McDaniel & Balistreri (2003), these methods for estimating trade elasticties were also inconsistent with the structure of AGE models, because, for example, they ignored supply-side considerations.

To estimate elasticities consistent with the general equilibrium structure of AGE models, Hummels (1999) exploits cross-sectional data, rather than relying on time-series data, as earlier studies did. Hummels estimates trade costs using both data on freight rates and model-implied estimates from a gravity equation, and then estimates elasticities using variation in imports and trade costs across countries. This methodology led to higher elasticities than previous studies. These higher elasticities were replicated in a study by Erkel-Rousse & Mirza (2002), who employ

an instrumental variable approach with time-series data to account for both demand- and supply-side considerations. These studies highlight the importance of estimating elasticities in a method consistent with the model with which the elasticities were to be used with.

Ruhl (2008) emphasizes the importance of making sure the elasticities correspond to the shocks being considered in the literature. Ruhl shows that temporary business cycle shocks imply lower elasticities than permanent shocks to trade costs, so one cannot use short-run elasticities to estimate the impact of trade reforms. Taken together, these and related findings emphasize the importance of taking seriously the relationships among shocks, elasticities, and the structures of models. As users of AGE models became aware of these concerns, they updated their methodologies for estimating trade elasticities. Hertel et al. (2007) provide estimated elasticities for the GTAP model by building off the work of Hummels (1999) and exploiting cross-sectional data on tariffs, trade flows, and other proxies for trade costs. Even with these improvements and newly estimated elasticities, however, our evaluation of the GTAP models shows that the performance of the models still has shortcomings.

The new trade models are promising because they bring along with them alternative interpretations for the trade elasticities. Chaney (2008) shows that the trade elasticity for the Melitz (2003) model of international trade, where the efficiency distribution across firms is Pareto, is pinned down by the tail parameter of the Pareto distribution. Similarly, Eaton & Kortum (2002) show that the trade elasticity in their model is given by the dispersion parameter from a Fréchet distribution for product-level productivities. These advances are exciting because they create the possibility of calibrating trade elasticities using methods that do not rely on cross-sectional or time-series variation in industry-level trade flows and tariffs. That firm- or product-level data can be useful in calibrating elasticities is supported by our evaluation comparison of the LTP predictions with those of the GTAP model in the previous section.

In practice, however, elasticity estimates based on these new trade models [e.g., those by Romalis (2007) and Caliendo & Parro (2015)], are still estimated using industry-level gravity equations that relate changes in trade flows to changes in tariffs. These estimates often use sophisticated differencing to avoid requiring data on domestic production and to eliminate the influence of symmetric trade barriers. As Simonovska & Waugh (2014) point out, however, these estimates do not depend on the specific structure of the model, and these gravity equations give equivalent estimates for all gravity estimates (i.e., they likely suffer from the same problems as the

GTAP elasticities). Simonovska & Waugh also show, however, that when micro-level price variation is combined with data on trade flows, the new trade models imply substantially different trade elasticities from the earlier Armington models and from each other. For a given trade elasticity, however, the behavior of the models is similar. This is why the calibration techniques from these models can be useful for improving the performance of AGE models. Feenstra et al. (2014) further investigate the tension between micro- and macro-estimates of trade elasticities and find significant differences between the estimates yielded by the two methods for roughly half of products, even when the estimates are carried out at the same level of disaggregation.

Related to our suggestion that the trade elasticity might depend on the product-level composition of trade flows, recent studies in the theoretical trade literature have also questioned the assumption of a constant trade elasticity. Fieler (2011) points out that different products have different trade elasticities, so the overall trade elasticity should change with the trade composition. Similarly, Simonovska (2015) examines data from an online retailer and provides evidence of non-homothetic consumer preferences, implying that a country's trade elasticity varies with its income level. Along a similar line, Jung et al. (2015) show that to properly match salient features of the data, we have to escape from the notion of having a constant trade elasticity. Our hope is that these insights on nonconstant elasticities will be helpful for figuring out ways to estimate trade elasticities that differ both by industry and by bilateral-country pair.

5.2. Evaluation of the Caliendo-Parro AGE Model

Although multi-industry, multi-country models featuring IO linkages across industries are not incompatible with the recent advances in the theoretical trade literature, relatively little work has been done to merge the two literatures. Recent exceptions are studies by Caliendo & Parro (2015) (henceforth we refer to their study as CP), Heerman et al. (2015), and Shikher (2012). These papers embed the framework of Eaton & Kortum (2002) into a multi-industry environment featuring IO linkages. These studies depart from standard AGE models in that trade costs, as opposed to variation in preferences or production functions, are the driving factor behind home bias and expenditure shares. One of the key innovations of CP has to do with their calibration of trade elasticities, which they are able to estimate at the industry level using asymmetries in tariffs, trade costs, and trade flows across countries. These elasticities are essential for determining the counterfactual trade response to changes in trade costs and tariffs and for translating this response

into welfare predictions. A benefit of merging AGE models with the Eaton & Kortum (2002) framework is that the models deliver recognizable gravity-type equations and transparent mappings to welfare predictions, which helps circumvent the criticism that AGE models act like black boxes.

One question is whether these new models outperform the standard AGE models, such as GTAP and the models originally used to predict NAFTA. To answer this question, we can look at the model predictions from CP, who calibrate their model using pre-NAFTA data on trade flows, production and IO linkages, and tariffs for the United States, Canada, Mexico, and 28 other countries. They then use their models to generate counterfactual predictions resulting from tariffs changing to their post-NAFTA levels. Although the focus of their study is on using their model to disentangle the welfare implications of NAFTA, we can also evaluate the accuracy of the model in matching actual changes in trade flows following the implementation of NAFTA.

We repeat the evaluation exercise from Section 4.2, computing the industry-weighted correlation between changes in observed industry-level trade flows for several variations of the CP model and for the share of LTP in each industry. Compared with our evaluation of GTAP models, there are a few changes. The industries used in the CP model differ from the industries defined in the GTAP. CP have 20 traded and 20 nontraded industries in their model, and we focus on the 20 traded industries, because we are evaluating only the accuracy in predicting changes in trade flows. CP provide a full description of the industries and a concordance between their industries and two-digit ISIC Rev. 3 industry codes in their paper. To compute actual changes in trade flows, we download trade data at the six-digit HS1988/1992 level, which we map into two-digit ISIC Rev. 3 industry codes using a concordance from the World Bank's World Integrated Trade Solution (WITS) database and then into our final industries using the concordance from CP. Unlike the GTAP simulations, which produce percentage changes, CP considers changes to be in terms of log differences (the results are similar regardless of whether log differences or percentage changes are used; however, we use log differences to be consistent with the original CP framework). Note also that unlike the GTAP models, CP considers NAFTA as the policy reform when computing their counterfactuals. Therefore, instead of setting up and solving an equivalent model, we are able to use the predicted changes generated by their own code, available in their data appendix (details in our own online data appendix).

To evaluate how well the CP predictions perform, we now compute actual changes in exports, $z_{ijk}^{t,t'}$, from country i to country j for each industry k between $t = 1991$ and $t' = 2006$ using the log approximation of equation (2):

$$z_{ijk}^{t,t'} = 100 \times \left(\log\left(x_{ijk}^{t'} \right) - \log\left(x_{ijk}^{t} \right) \right), \tag{7}$$

where x_{ijk}^{t} is exports from country i to country j in industry k in period t, reported in current price USD. We provide an appendix that shows that our results are robust to calculated changes in exports using either method and to excluding or including outliers. For the least traded exercise, the share of LTP in each industry is computed in the same way as in Section 4.2, where, for each importer-exporter pair, we sort products by their average trade value between 1991 and 1993 and then count products as least traded until they cumulatively account for exactly 10 percent of trade in 1991.

We report the results of these exercises in Table 2. The column titled "CP correlation with data" refers to the weighted correlation between actual changes in trade flows and the predicted changes of the full CP model, taking into account all tariff changes prior to 2006. This is computed using equations (4)–(6), with the predicted changes for each industry taken from CP and the actual changes computed using (7). The column titled "LTP correlation with data" provides the same comparison benchmark of the weighted correlation between actual changes in trade flows and the share of LTP in any industry, which we compute following the same methodology as we did for Table 1.

So does the CP framework outperform standard AGE models? The answer is: It depends. The average correlation across country pairs is near zero and slightly lower than the GTAP correlations in Table 1; therefore, it may appear that there are little to no gains from incorporating recent advances into AGE models. In contrast to the GTAP results in Table 1, however, the CP framework outperforms the LTP methodology for half of the country pairs, whereas the GTAP framework is uniformly beaten by the LTP methodology. Similarly, the correlations between actual and CP predicted exports from the United States to Canada and Mexico are high and significantly higher than any of the correlations between actual changes and predicted changes for the GTAP model. This can be regarded as a partial success of the CP framework and the improvements that are possible when incorporating insights from the new trade models. Although

it may initially appear unfair to compare the accuracy of the CP predictions, which take into account all tariff changes, with the GTAP predictions, which take into account only changes in tariffs between the member countries of the free trade agreement, CP also computes counterfactuals taking into account only NAFTA tariff changes. The column titled "CP correlation with data (only NAFTA tariffs)" in Table 2 shows that the results are nearly unchanged if tariff changes in non-NAFTA countries are disregarded.

Table 2: Comparisons of CP and LTP predictions for NAFTA with data

Exporter	Importer	CP correlation with data	CP correlation with data (only NAFTA tariffs)	CP correlation with data (no IO structure)	LTP correlation with data
Canada	Mexico	−0.46	−0.46	−0.50	0.27
Canada	United States	0.36	0.32	0.36	0.19
Mexico	Canada	−0.68	−0.66	−0.71	0.83
Mexico	United States	−0.17	−0.12	−0.21	0.33
United States	Canada	0.35	0.05	0.14	0.28
United States	Mexico	0.54	0.53	0.64	0.16
Simple average		−0.01	−0.06	−0.05	0.33

Although the CP model shows considerable success for some country pairs, it also shows considerable failure in producing accurate predictions for trade between other country pairs, particularly trade between Canada and Mexico. Why does their model perform so poorly in these cases? We hypothesize that it is because the CP methodology lacks the LTP margin. Caliendo & Parro (2015) calibrate the crucial Fréchet parameters in their model using the same methodology — one that relies on an industry-level gravity equation, albeit in a novel form — that other researchers use to calibrate Armington elasticities. In other words, they do not exploit the features that make their model different from an Armington model like the GTAP model. We should stress that the extensive margin at the firm level in the Eaton-Kortum (2002) and Melitz (2003) models as they are usually parameterized has little or nothing to do with the extensive margin at the product-level studied by Kehoe et al. (2015).

To understand this point, note that trade between Canada and Mexico is exactly where the LTP methodology performs best. Indeed, we find that, for exports from Mexico to Canada, LTP grew by 206.0 percent more than GDP between 1992 and 2006, whereas non–LTP grew only 55.8 percent. Why do least traded products grow so much more than non–least traded products? The baseline models of Eaton & Kortum (2002) and Melitz (2003) are incapable of reproducing this observation, at least as they are usually parameterized. One model that is capable of capturing the

Kehoe & Ruhl (2013) extensive margin is that of Arkolakis (2010), in which acquiring new customers is costly, and nonlinear marketing costs can explain why small firms grow faster than large firms. Potentially in agreement with this theory, Ruhl & Willis (forthcoming) show that exporters tend to start small and grow over time, whereas Schmeiser (2012) shows that the expansion of exporters to new markets occurs slowly.

The column titled "CP correlation with data (no IO structure)" in Table 2 reports the weighted correlation between actual changes in trade flows and the counterfactuals Caliendo & Parro (2015) produce using their framework for NAFTA tariff changes only and discarding the input-output structure of their model. It is disconcerting that the CP framework performs nearly identically (and actually slightly better on average) when the IO structure of the model is left out. Above, we argue that taking into account linkages across industries is essential if we want to understand the industry-level impact of trade reforms. The fact that the IO structure does not improve the performance of the CP model in general suggests that economists need to think more carefully about the way in which input-output relationships are being built into AGE trade models and to what extent these issues could be resolved with elasticity estimates that vary bilaterally.

Overall, the results in Table 2 suggest that merging recent trade advances with AGE models shows promise in improving the performance of these models but that further improvements are needed. Some of these improvements may be relatively easy to attain, for example, calibrating elasticities in a way that delivers different estimates for different model structures, as do Simonovska & Waugh (2014). The fact that the LTP methodology performs best precisely when the CP model performs worse and vice versa hints at a likely pathway for improvement. Although further research will need to confirm whether comparable findings hold for other trade liberalizations, the fact that the LTP methodology performs better on average than the CP model and uniformly outperforms the GTAP models in Section 4.2 suggests that the gains from incorporating the LTP margin in AGE models are potentially large. Our benchmark predictions suggest that estimating industry-level elasticities that are country-pair specific and depend on the product composition of industries within each country is a promising avenue for future research.

6. Quantifying the Gains from Trade and Future Directions for AGE Models

As Shoven & Whalley (1984) put it, AGE models aim "to convert the Walrasian general-equilibrium structure […] from an abstract representation of an economy into realistic models of actual economies" to "use these models to evaluate policy options by specifying production and

demand parameters and incorporating data reflective of real economies." Ultimately, AGE models are used to help researchers communicate the welfare implications of policy reforms. As we mention above, one of the biggest breakthroughs in the recent trade literature is the formula derived by Arkolakis et al. (2012) — which we refer to as the ACR formula — for computing welfare gains from trade. They show that for a wide class of models, capturing the welfare impact of a trade reform requires only the change in trade that the reform produces and the trade elasticity. The basic formula derived in their study is

$$\hat{w} = \hat{\lambda}^{1/\varepsilon} \ , \tag{8}$$

where $\hat{\lambda}$ is the change in domestic expenditure share (or self-trade share) produced by the reform, ε is the trade elasticity (the degree to which imports respond to changes in trade costs), and $\hat{w}$ is the resulting change in welfare implied by this class of models.

This result has been widely used in recent studies because it allows for changes in welfare to be computed easily, especially when comparing gains from moving away from autarky when the self-trade share is one. Costinot & Rodríguez-Clare (2014) use the ACR formula to quantitatively evaluate the welfare implications of globalization. Relevant to evaluations of NAFTA, this basic formula computes changes in welfare resulting from changes in iceberg trade costs rather than in tariffs. Felbermayr et al. (2015) argue that this distinction has welfare implications because tariff revenue is redistributed back to consumers. Goldberg & Pavcnik (2016) further point out that, although tariffs are now relatively low for most of the world, there is significant room for trade policy to address trade costs associated with nontariff barriers that may more closely resemble iceberg trade costs. The insights from the ACR formula have also been expanded, for instance, by Adao et al. (2015), who show how welfare gains from trade can be measured in the absence of parametric specifications of functional forms, and by Brooks & Pujolàs (2014), who develop a generalized version of the ACR formula for demand systems featuring a nonconstant trade elasticity. Even before these developments, however, the literature was shifting away from using AGE models to evaluate the gains from trade, as evidenced by Alvarez & Lucas (2007), who study the welfare implications of the Eaton & Kortum (2002) model.

Why, then, do we still need multi-industry AGE models? First, many policy questions are concern more than just simple welfare or changes in overall trade flows. Second, even if we were only interested in welfare, and the models were able to correctly forecast changes in aggregate

bilateral trade, the composition of that trade matters for welfare gains because the aggregate trade elasticity depends on the industry-level composition of trade. Ossa (2015) shows that accounting for differences in elasticities across industries greatly increases estimates for welfare gains of trade compared with single-sector models that do not differentiate between industries. Countries experience large welfare gains from importing goods in low-elasticity industries, such as automobiles, in which they are not efficient at producing domestically. French (2016) further shows how the pattern of comparative advantage across industries affects welfare gains as well as the aggregate impact of trade barriers, the insight being that multi-industry models are needed even if the goal is only to capture changes in aggregate trade flows and not in welfare or disaggregated trade flows.

Another distinguishing feature of AGE models is their focus on the IO structure of economies. Taking the IO structure into account is essential for understanding the nature and impact of trade flows because trade in intermediate goods makes up a large fraction of international trade. As Yi (2003) and Ramanarayanan (2012) show, the impact of tariffs and other trade barriers is amplified when there is trade in intermediate goods, as the trade costs apply both directly to trade in final goods and indirectly through their embodied impact on intermediate goods. This result is similar to the concept of double marginalization in the industrial organization literature. Although trade in intermediate goods can be accounted for in single-sector models, as it is by Eaton & Kortum (2002), incorporating IO linkages allows for the richer interactions that are necessary to capture how tariff reductions in one industry affect trade and production in other industries. In addition, the studies mentioned in this section suggest that AGE features, such as IO linkages across industries and industry-level elasticity estimates, are undoubtedly important to the welfare implications of trade.

Before concluding this study, it is important to bring attention to the fact that, as is the case in much of the literature, our focus is on static AGE models. Although the focus of this paper is therefore relatively narrow, the use of dynamic AGE models in policy-related applications (e.g., Diao et al. (1998), Ianchovichina (2012)) has increased. Despite this progress, however, many features and implications of dynamic trade models have also yet to be fully explored in AGE settings.

The potential impact of building dynamics into trade models is often nontrivial. As Bajona & Kehoe (2010) show, a standard trade model such as the Heckscher-Ohlin model, when built into

a dynamic framework, implies that standard results in closed-economy dynamic models, such as long-run convergence of income, disappear. Baldwin (1992) and Anderson et al. (2015) show that there are large dynamic gains from trade because it encourages capital accumulation and human capital accumulation. Building on this argument, Brooks & Pujolàs (2016) show that gains from trade driven by capital accumulation can be large, even though, in the transition, they are potentially negative. Studying the impact that trade reforms have on the creation of new firms, Alessandria et al. (2014) and Alessandria & Choi (2015) show that gains during the transition can be even larger. Although we need to first fix the problems with static AGE models, incorporating these features about dynamics into AGE models should eventually provide us with better estimates of the impact of trade liberalization.

7. Final Remarks

Throughout this paper, we have highlighted shortcomings of AGE models that must be addressed if they are to continue to be used to evaluate the impact of a trade reform. We conclude with some conjectures about the best path for research going forward. First, despite the performance issues we identify with AGE models, they are still the only models currently being used that address the industry-level impact of changes in trade policy. We conjecture that, when these shortcomings are addressed, they will remain the tool of choice for evaluating trade policy. The issues we identify primarily have to do with important elements that are missing from AGE models, such as the increased growth from LTP and not incorporating more nuanced measures of gains from trade. The solution, therefore, should be to improve AGE models by embedding these elements, not to shift to alternative methodologies that throw away the important features AGE models have already successfully included. For instance, although our benchmark LTP predictions were useful in showing the potential gains from introducing the LTP margin into AGE models, they cannot legitimately be used to inform policy makers because they lack essential elements and cannot, for example, be used to inform on how trade will affect wages, the impact of linkages across industries, and spillover effects to countries not included in the liberalization.

Instead of encouraging disillusionment with AGE models, our goal is to spur their improvement. For example, it appears likely that AGE models can be improved by embedding the LTP margin. One of the possible methods for doing so could be to introduce a marketing cost on firms, in the spirit of Arkolakis (2010), where firms have the potential to grow quickly after entering markets through additional marketing that leads more consumers to become aware of their

products. In this model, the LTP margin appears naturally as the marketing costs necessary to reach an additional consumer increase because fewer consumers remain unaware of the firm's product. We find it likely that there are other avenues by which the LTP margin can emerge as well. The next step, therefore, is to find them and evaluate which ones perform best at matching the observed patterns in the data. By repeating this process of identifying shortcomings and addressing them, we expect that researchers will be able to deliver a new generation of AGE models that are better able to predict the industry impact of trade reforms.

Although we are optimistic about the future of AGE models, it is also important to comment on the past and present. Readers should not conclude from our results that, because AGE models do not accurately predict which industries will grow the most and the least following trade reforms, that economists are ill informed on the welfare impact of trade. On the contrary, the overwhelming consensus is that the welfare gains from trade are positive: Free trade reduces the prices that consumers face, increases the varieties available, and expands the sales of productive firms. Even in cases in which trade does not increase GDP, it can still be expected to increase welfare (Kehoe & Ruhl (2010)). In fact, standard AGE models of trade are likely too conservative on the predicted welfare gains from liberalization. To illustrate this, we calculate the additional trade that occurred for each of the countries in our exercises because of the higher growth rates of LTP. For the country pairs we used in our evaluation of the GTAP predictions, we find that, on average, overall bilateral trade flows grew 18.9 percent more than they would have if LTP goods had grown at the rates of non-LTP. For NAFTA countries, the number is even higher, with an average increase of 27.2 percent. These results indicate that a model that is unable to capture the LTP margin is likely to underestimate the increase in welfare gains from trade resulting from a trade liberalization. This means, of course, that a substantial part of the benefit from trade agreements — for example the gains from NAFTA for the United States — has occurred in industries in which such gains were not expected or predicted when the liberalizations were implemented. Given the frequent disconnect between economists and laypeople on the impact of trade policy, it is necessary for us to clearly and correctly identify and communicate the gains from trade to policy makers and society. Improving AGE models will allow economists to do this.

Literature Cited

Adao R, Costinot A, Donaldson D. 2015. *Nonparametric counterfactual predictions in neoclassical models of international trade*. NBER Work. Pap. 21401

Alessandria G, Choi H. 2015. *The dynamics of the trade balance and the real exchange rate: the J curve and trade costs?* SED 2015 Meet. Pap. 1413

Alessandria G, Choi H, Ruhl KJ. 2014. *Trade adjustment dynamics and the welfare gains from trade*. NBER Work. Pap. 20663

Allayannis G, Ofek E. 2001. Exchange rate exposure, hedging, and the use of foreign currency derivatives. *J. Int. Money Finance* 20:273–96

Allen T. 2014. Information frictions in trade. *Econometrica* 82:2041–83

Alvarez F, Lucas Jr. RE. 2007. General equilibrium analysis of the Eaton–Kortum model of international trade. *J. Monet. Econ.* 54:1726–68

Anderson JE, Larch M, Yotov YV. 2015. *Growth and trade with frictions: a structural estimation framework*. NBER Work. Pap. 21377

Anderson JE, van Wincoop E. 2003. Gravity with gravitas: a solution to the border puzzle. *Am. Econ. Rev.* 93:170–92

Anderson JE, van Wincoop E. 2004. Trade costs. *J. Econ. Lit.* 42:691–751

Arkolakis C. 2010. Market penetration costs and the new consumers margin in international trade. *J. Polit. Econ.* 118:1151–99

Arkolakis C, Costinot A, Rodríguez-clare A. 2012. New trade models, same old gains? *Am. Econ. Rev.* 102:94–130

Armington PS. 1969. A theory of demand for products distinguished by place of production. *IMF Staff Pap.* 16:159–78

Asturias J. 2016. *Endogenous transportation costs*. Work. Pap., Georgetown Univ.

Bagwell K, Staiger RW. 1999. An economic theory of GATT. *Am. Econ. Rev.* 89:215–48

Baier SL, Bergstrand JH. 2007. Do free trade agreements actually increase members' international trade? *J. Int. Econ.* 71:72–95

Bajona C, Kehoe TJ. 2010. Trade, growth, and convergence in a dynamic Heckscher-Ohlin model. *Rev. Econ. Dyn.* 13:487–513

Baldwin RE. 1992. Measurable dynamic gains from trade. *J. Polit. Econ.* 100:162–74

Böhringer C, Löschel A. 2006. Computable general equilibrium models for sustainability impact assessment: status quo and prospects. *Ecol. Econ.* 60:49–64

Broda C, Limao N, Weinstein DE. 2008. Optimal tariffs and market power: the evidence. *Am. Econ. Rev.* 98:2032–65

Brooks WJ, Pujolàs PS. 2014. *Nonlinear gravity*. McMaster Univ. Work. Pap. 2014-15

Brooks WJ, Pujolàs PS. 2016. *Capital accumulation and the welfare gains from trade*. McMaster Univ. Work. Pap. 2016-03

Brown DK, Deardorff AV, Stern RM. 1992. A North American free trade agreement: analytical issues and a computational assessment. *World Econ.* 15:11–30

Brown DK, Stern RM. 1989. U.S.-Canada bilateral tariff elimination: the role of product differentiation and market structure. In *Trade Policies for International Competitiveness*, ed. RC Feenstra, pp. 217–54. Chicago: University of Chicago Press

Burfisher ME, Dyck J, Meade B, Mitchell L, Wainio JT, et al. 2014. *Agriculture in the Trans-Pacific Partnership*. USDA-ERS ERR 176

Burfisher ME, Robinson S, Thierfelder K. 2001. The impact of NAFTA on the United States. *J. Econ. Perspect.* 15:125–44

Burniaux J-M, Truong TP. 2002. *GTAP-E: an energy-environmental version of the GTAP model*. GTAP Tech. Pap. 16

Caliendo L, Parro F. 2015. Estimates of the trade and welfare effects of NAFTA. *Rev. Econ. Stud.* 82:1–44

Chaney T. 2008. Distorted gravity: the intensive and extensive margins of international trade. *Am. Econ. Rev.* 98:1707–21

Comerford D, Rodriguez-Mora JV. 2015. *The gains from economic integration*. SED 2015 Meet. Pap. 569

Corsetti G, Martin P, Pesenti P. 2007. Productivity, terms of trade and the "home market effect." *J. Int. Econ.* 73:99–127

Costinot A, Donaldson D, Vogel J, Werning I. 2015. Comparative advantage and optimal trade policy. *Q. J. Econ.* 130:659–702

Costinot A, Rodríguez-Clare A. 2014. Trade theory with numbers: quantifying the consequences of globalization. In *Handbook of International Economics*, Vol. 4, ed. E Helpman, K Rogoff, G Gopinath, pp. 197–261. Elsevier

Cox D, Harris RG. 1992. North American free trade and its implications for Canada: results from a CGE model of North American trade. *World Econ.* 15:31–44

Dean JM, Signoret JE, Feinberg RM, Ludema RD, Ferrantino MJ. 2009. Estimating the price effects of non-tariff barriers. *BE J. Econ. Anal. Policy* 9.

Dekle R, Eaton J, Kortum S. 2008. Global rebalancing with gravity: measuring the burden of adjustment. *IMF Staff Pap.* 55:511–40

Demidova S, Rodriguez-Clare A. 2009. Trade policy under firm-level heterogeneity in a small economy. *J. Int. Econ.* 78:100–112

Dervis K, de Melo J, Robinson S. 1982. *General equilibrium models for development policy*. Cambridge, England: Cambridge University Press

Diao X, Roe TL, Yeldan E. 1998. A simple dynamic applied general equilibrium model of a small open economy: transitional dynamics and trade policy. *J. Econ. Dev.* 23:77–101

Dixit AK, Stiglitz JE. 1977. Monopolistic competition and optimum product diversity. *Am. Econ. Rev.* 67:297–308

Donaldson D. 2015. The gains from market integration. *Annu. Rev. Econ.* 7:619–47

Donaldson D. Forthcoming. Railroads of the Raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.*

Eaton J, Kortum S. 2002. Technology, geography, and trade. *Econometrica* 70:1741–79

Egger P. 2000. A note on the proper econometric specification of the gravity equation. *Econ. Lett.* 66:25–31

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

Estevadeordal A, Frantz B, Taylor AM. 2003. The rise and fall of world trade, 1870–1939. *Q. J. Econ.* 118:359–407

Feenstra RC, Luck PA, Obstfeld M, Russ KN. 2014. *In search of the Armington elasticity*. NBER Work. Pap. 20063

Feenstra RC, Yang T-H, Hamilton GG. 1999. Business groups and product variety in trade: evidence from South Korea, Taiwan and Japan. *J. Int. Econ.* 48:71–100

Felbermayr G, Jung B, Larch M. 2015. The welfare consequences of import tariffs: a quantitative perspective. *J. Int. Econ.* 97:295–309

Fieler AC. 2011. Nonhomotheticity and bilateral trade: evidence and a quantitative explanation. *Econometrica*. 79:1069–1101

Fox AK. 1999. *Evaluating the success of a CGE model of the Canada-U.S. Free Trade Agreement*. Unpublished Manuscript, Univ. Michigan

French S. 2016. The composition of trade flows and the aggregate effects of trade barriers. *J. Int. Econ.* 98:114–37

Goldberg PK, Pavcnik N. 2016. *The effects of trade policy*. NBER Work. Pap. 21957

Grossman GM, Krueger AB. 1994. Environmental impacts of a North American Free Trade Agreement. In *The Mexico-U.S. Free Trade Agreement*, ed. PM Garber, pp. 13–56. Cambridge, MA: MIT Press

Grubel HG, Lloyd PJ. 1971. The empirical measurement of intra-industry trade. *Econ. Rec.* 47:494–517

Hanson GH, Xiang C. 2004. The home-market effect and bilateral trade patterns. *Am. Econ. Rev.* 94:1108–29

Harris R. 1984. Applied general equilibrium analysis of small open economies with scale economies and imperfect competition. *Am. Econ. Rev.* 74:1016–32

Hazilla M, Kopp RJ. 1990. Social cost of environmental quality regulations: a general equilibrium analysis. *J. Polit. Econ.* 98:853–73

Head K, Mayer T. 2014. Gravity equations: workhorse, toolkit, and cookbook. In *Handbook of International Economics*, Vol. 4, ed. G Gopinath, E Helpman, K Rogoff, pp. 131–95. Amsterdam: Elsevier

Head K, Mayer T, Ries J. 2010. The erosion of colonial trade linkages after independence. *J. Int. Econ.* 81:1–14

Heerman KER, Arita S, Gopinath M. 2015. Asia-Pacific integration with China versus the United States: examining trade patterns under heterogeneous agricultural sectors. *Am. J. Agric. Econ.* 97:1324–44

Hertel T. 2013. Global applied general equilibrium analysis using the global trade analysis project framework. In *Handbook of Computable General Equilibrium Modeling*, Vol. 1, ed. PB Dixon, D Jorgenson, pp. 815–76. Amsterdam: Elsevier

Hertel T, Hummels D, Ivanic M, Keeney R. 2007. How confident can we be of CGE-based assessments of free trade agreements? *Econ. Model.* 24:611–35

Hillberry RH, McDaniel CA. 2002. A decomposition of North American trade growth since NAFTA. *Int. Econ. Rev.* USITC Pub. No. 3527:1–6

Hummels D. 1999. *Toward a geography of trade costs*. *1162*, Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University

Hummels D, Klenow PJ. 2005. The variety and quality of a nation's exports. *Am. Econ. Rev.* 95:704–23

Hummels D, Lugovskyy V, Skiba A. 2009. The trade reducing effects of market power in international shipping. *J. Dev. Econ.* 89:84–97

Hummels DL, Schaur G. 2013. Time as a trade barrier. *Am. Econ. Rev.* 103:2935–59

Ianchovichina E. 2012. *Dynamic modeling and applications for global economic analysis*. Cambridge, England: Cambridge University Press

Irarrazabal A, Moxnes A, Opromolla LD. 2015. The tip of the iceberg: a quantitative framework for estimating trade costs. *Rev. Econ. Stat.* 97:777–92

Jacks DS, Meissner CM, Novy D. 2008. Trade costs, 1870–2000. *Am. Econ. Rev.* 98:529–34

Jacks DS, Meissner CM, Novy D. 2011. Trade booms, trade busts, and trade costs. *J. Int. Econ.* 83:185–201

Jung JW, Simonovska I, Weinberger A. 2015. *Exporter heterogeneity and price discrimination: a quantitative view*. NBER Work. Pap. 21408

Kehoe PJ, Kehoe TJ. 1994a. A primer on static applied general equilibrium models. *Fed. Reserve Bank Minneap. Q. Rev.* 18:2–16

Kehoe PJ, Kehoe TJ. 1994b. Capturing NAFTA's impact with applied general equilibrium models. *Fed. Reserve Bank Minneap. Q. Rev.* 18:17–34

Kehoe TJ. 2005. An evaluation of the performance of applied general equilibrium models on the impact of NAFTA. In *Frontiers in Applied General Equilibrium Modeling*, ed. TJ Kehoe, TN Srinivasan, J Whalley, pp. 341–77. Cambridge, England: Cambridge University Press

Kehoe TJ, Noyola PJ, Manresa A, Polo C, Sancho F. 1988. A general equilibrium analysis of the 1986 tax reform in Spain. *Eur. Econ. Rev.* 32:334–42

Kehoe TJ, Polo C, Sancho F. 1995. An evaluation of the performance of an applied general equilibrium model of the Spanish economy. *Econ. Theory* 6:115–41

Kehoe TJ, Prescott EC. 1995. Introduction to the symposium: the discipline of applied general equilibrium. *Econ. Theory* 6:1–11

Kehoe TJ, Rossbach J, Ruhl KJ. 2015. Using the new products margin to predict the industry-level impact of trade reform. *J. Int. Econ.* 96:289–97

Kehoe TJ, Ruhl KJ. 2010. Why have economic reforms in Mexico not generated growth? *J. Econ. Lit.* 48:1005–27

Kehoe TJ, Ruhl KJ. 2013. How important is the new goods margin in international trade? *J. Polit. Econ.* 121:358–92

Kehoe TJ, Ruhl KJ, Steinberg JB. 2013. *Global imbalances and structural change in the United States*. *19339*, NBER Work. Pap. 19339

Kleinert J, Spies J. 2011. *Endogenous transport costs in international trade*. IAW Disc. Pap. 74

Kohn D, Leibovici F, Szkup M. 2016. Financial frictions and new exporter dynamics. *Int. Econ. Rev.* 57:453–86

Krugman P. 1980. Scale economies, product differentiation, and the pattern of trade. *Am. Econ. Rev.* 70:950–59

Leibovici F. 2015. *Financial development and international trade*. York Univ. Work. Pap. 2015–3

Li C, Whalley J. 2014. China and the Trans-Pacific Partnership: a numerical simulation assessment of the effects involved. *World Econ.* 37:169–92

Limão N, Venables AJ. 2001. Infrastructure, geographical disadvantage, transport costs, and trade. *World Bank Econ. Rev.* 15:451–79

Maggi G, Rodríguez-Clare A. 2007. A political-economy theory of trade agreements. *Am. Econ. Rev.* 97:1374–1406

Manova K. 2013. Credit constraints, heterogeneous firms, and international trade. *Rev. Econ. Stud.* 80:711–44

Markusen JR, Hunter L, Rutherford TF. 1995. Trade liberalization in a multinational-dominated industry. *J. Int. Econ.* 38:95–117

McCallum J. 1995. National borders matter: Canada-U.S. regional trade patterns. *Am. Econ. Rev.* 85:615–23

McDaniel CA, Balistreri EJ. 2003. A review of Armington trade substitution elasticities. *Econ. Int.* 23:301–13

McKitrick RR. 1998. The econometric critique of computable general equilibrium modeling: the role of functional forms. *Econ. Model.* 15:543–73

Melitz MJ. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71:1695–1725

Narayanan GB, Ciuriak D, Singh HV. 2016. Quantifying the mega-regional trade agreements: a review of the models. In *TPP and India: Implications of Mega-Regionals for Developing Economies*, ed. HV Singh, pp. 93–131. New Delhi, India: Wisdom Tree

New Zealand Ministry of Foreign Affairs and Trade and China Ministry of Commerce. 2004. *A joint study report on a free trade agreement between China and New Zealand*

Obstfeld M, Rogoff K. 2001. The six major puzzles in international macroeconomics: is there a common cause? In *NBER Macroeconomics Annual 2000*, Vol. 15, ed. BS Bernanke, K Rogoff, pp. 339–412. Cambridge, MA: MIT Press

Opp MM. 2010. Tariff wars in the Ricardian model with a continuum of goods. *J. Int. Econ.* 80:212–25

Ossa R. 2015. Why trade matters after all. *J. Int. Econ.* 97:266–77

Ottaviano G, Pessoa JP, Sampson T, van Reenen J. 2014. *The costs and benefits of leaving the EU*. CFS Work. Pap. 472

Pavcnik N. 2002. Trade liberalization, exit, and productivity improvements: evidence from Chilean plants. *Rev. Econ. Stud.* 69:245–76

Peterson EB, Schleich J, Duscha V. 2011. Environmental and economic effects of the Copenhagen pledges and more ambitious emission reduction targets. *Energy Policy* 39:3697–3708

Ramanarayanan A. 2012. *Imported inputs and the gains from trade*. SED 2012 Meet. Pap. 612

Roberts MJ, Tybout JR. 1997. The decision to export in Colombia: an empirical model of entry with sunk costs. *Am. Econ. Rev.* 87:545–64

Romalis J. 2007. NAFTA's and CUSFTA's impact on international trade. *Rev. Econ. Stat.* 89:416–35

Ruhl KJ. 2008. *The international elasticity puzzle*. Work. Pap., Penn State Univ.

Ruhl KJ, Willis JL. Forthcoming. New exporter dynamics. *Int. Econ. Rev.*

Samuelson PA. 1954. The transfer problem and transport costs, ii: analysis of effects of trade impediments. *Econ. J.* 64:264–89

Santos Silva JMC, Tenreyro S. 2006. The log of gravity. *Rev. Econ. Stat.* 88:641–58

Schmeiser KN. 2012. Learning to export: export growth and the destination decision of firms. *J. Int. Econ.* 87:89–97

Shapiro JS. 2016. Trade costs, CO2, and the environment. *Am. Econ. J. Econ. Policy*. 8:220–54

Shiells CR, Stern RM, Deardorff AV. 1986. Estimates of the elasticities of substitution between imports and home goods for the United States. *Weltwirtschaftliches Arch.* 122:497–519

Shikher S. 2012. Predicting the effects of NAFTA: now we can do it better! *J. Int. Glob. Econ. Stud.* 5:32–59

Shoven JB, Whalley J. 1984. Applied general-equilibrium models of taxation and international trade: an introduction and survey. *J. Econ. Lit.* 22:1007–51

Shoven JB, Whalley J. 1992. *Applying general equilibrium*. Cambridge, England: Cambridge University Press

Simonovska I. 2015. Income differences and prices of tradables: insights from an online retailer. *Rev. Econ. Stud.* 82:1612–56

Simonovska I, Waugh ME. 2014. *Trade models, trade elasticities, and the gains from trade*. NBER Work. Pap. 20495

Smith A, Venables AJ. 1988. Completing the internal market in the European community: some industry simulations. *Eur. Econ. Rev.* 32:1501–25

Sobarzo HE. 1995. A general equilibrium analysis of the gains from NAFTA for the Mexican economy. In *Modeling North American Economic Integration*, ed. PJ Kehoe, TJ Kehoe, pp. 91–115. Boston: Kluwer Academic

Stern RM, Francis J, Schumacher B. 1976. *Price elasticities in international trade: an annotated bibliography*. Macmillan for the Trade Policy Research Centre

Timmer MP, Dietzenbacher E, Los B, Stehrer R, de Vries GJ. 2015. An illustrated user guide to the World Input–Output Database: the case of global automotive production. *Rev. Int. Econ.* 23:575–605

Trefler D. 2004. The long and short of the Canada-U.S. Free Trade Agreement. *Am. Econ. Rev.* 94:870–95

Tybout J, de Melo J, Corbo V. 1991. The effects of trade reforms on scale and technical efficiency: new evidence from Chile. *J. Int. Econ.* 31:231–50

US International Trade Commission. 2003. *U.S.-Chile Free Trade Agreement: potential economy-wide and selected sectoral effects*. USITC Pub. 3605

Waugh ME. 2010. International trade and income differences. *Am. Econ. Rev.* 100:2093–2124

Whalley J. 1985. *Trade liberalization among major world trading areas*. Cambridge, MA: MIT Press

Wolf HC. 2000. Intranational home bias in trade. *Rev. Econ. Stat.* 82:555–63

Yi K-M. 2003. Can vertical specialization explain the growth of world trade? *J. Polit. Econ.* 111:52–102

Yi K-M. 2010. Can multistage production explain the home bias in trade? *Am. Econ. Rev.* 100:364–93

A. Appendix

The following is a selection from our online appendix, which explores the robustness of our results.

A.1. Log Differences versus Percentage Changes

When evaluating the accuracy of the GTAP predictions in Table 1, we define growth in terms of percentage changes. When evaluating the accuracy of the CP predictions in Table 2, we define growth in terms of log differences. We choose to use a different definition for growth between the two tables for two reasons. First, we choose each definition to be consistent with the definition used by the original studies. GTAP studies use percentage changes, whereas CP uses log differences. Second, the choice shows that the overall shortcomings of AGE models do not depend on exactly how we define growth (percentage changes or log differences).

To explore the robustness of our results, we show that Tables 1 and 2 are similar overall when the definitions of growth are switched between them. Note that we can move between percentage changes and log differences with the following equations:

$$\text{log diff} = 100 * \log\left(\frac{\text{percentage change}}{100} + 1\right), \tag{.9}$$

$$\text{percentage change} = 100 * \left(\exp\left(\frac{\text{log diff}}{100}\right) - 1\right). \tag{.10}$$

Tables A.1 and A.2 show the results of this exercise.

Table A.1: Comparisons of GTAP and LTP predictions of recent trade liberalizations with data (log differences)

Exporter	Importer	GTAP correlation with data	LTP correlation with data
United States	Australia	0.63	0.49
Australia	United States	–0.18	0.45
United States	Chile	0.32	0.56
Chile	United States	0.15	0.44
China	Chile	0.51	0.78
Chile	China	–0.09	–0.08
China	New Zealand	–0.38	0.56
New Zealand	China	–0.17	0.30
Simple average		0.10	0.44

Overall, the LTP correlation in Table A.1 stays almost the same as that in Table 1. The biggest changes are that GTAP appears to perform better for Chinese exports to Chile and for U.S.

exports to Australia when expressed in log differences, whereas Chilean exports to China perform worse. Overall, the average LTP correlation is similar across the two tables (0.44 in Table A.1 compared with 0.49 in Table 1). The average GTAP correlation is slightly higher with log differences (0.10 compared with –0.00). It still performs significantly worse than the LTP methodology with both definitions, however.

Table A.2: Comparisons of CP and LTP predictions of NAFTA with data (percentage changes)

Exporter	Importer	CP correlation with data	LTP correlation with data
Canada	Mexico	−0.47	0.34
Canada	United States	0.63	0.05
Mexico	Canada	−0.46	0.76
Mexico	United States	−0.11	0.14
United States	Canada	0.38	0.20
United States	Mexico	0.96	−0.02
Simple average		0.16	0.28

The results in Table A.2 are similar to those in Table 2, with the LTP methodology outperforming the CP predictions. By far the biggest change is for U.S. exports to Mexico, which exhibit a negative correlation for LTP and a near perfect correlation for CP. This result is driven entirely by growth in exports in the petroleum industry, which grew by over 1500 percent between 1991 and 2006 (the second highest growth was in chemicals, which grew by 200 percent). Despite this huge growth, petroleum exports actually grew less than predicted. The CP model predicted an increase in growth of nearly 3000 percent (the second highest predicted growth was in electrical machinery, at nearly 200 percent). The CP model predicts this large increase because of a much higher estimated trade elasticity for the petroleum industry than for any other industry (over 50; no other industry was over 20). If the petroleum industry is dropped when evaluating the performance of each methodology, then the results are as shown in Table A.3.

Table A.3: Comparisons of CP and LTP predictions of NAFTA with data (percentage changes, U.S. petroleum exports to Mexico excluded)

Exporter	Importer	CP correlation with data	LTP correlation with data
Canada	Mexico	−0.47	0.34
Canada	United States	0.63	0.05
Mexico	Canada	−0.46	0.76
Mexico	United States	−0.11	0.14
United States	Canada	0.38	0.20
United States	Mexico	−0.37	0.45
Simple average		−0.06	0.35

As we can see, simply excluding the petroleum industry changes the correlation between the predictions of the CP model and actual growth for U.S. exports to Mexico from 0.96 to −0.37. One of the effects of log differences is that it lessens the overall influence of extreme growth rates for individual industries compared with using percentage changes.

A.2. Effects of Outlier Observations

When evaluating the GTAP predictions, we have excluded two industries: oil exports from the United States to Chile and cattle meat exports from Australia to the United States. Although excluding these industries does affect the correlations for each of those importer-exporter pairs, removing them has little effect on the overall performance of the GTAP predictions and LTP methodology. Table A.4 shows that LTP methodology performs substantially better than the GTAP predictions even when these outlier industries are not excluded from the analysis.

Table A.4: Comparisons of GTAP and LTP predictions of recent trade liberalizations with data (percentage changes, outliers not excluded)

Exporter	Importer	GTAP correlation with data	LTP correlation with data
United States	Australia	0.27	0.55
Australia	United States	0.49	0.08
United States	Chile	0.05	–0.05
Chile	United States	0.03	0.48
China	Chile	0.14	0.61
Chile	China	0.04	0.07
China	New Zealand	–0.36	0.61
New Zealand	China	–0.09	0.48
Simple average		0.07	0.35

Note that including cattle meat exports from Australia to the United States makes the GTAP predictions appear to perform much better for that pair. It is, however, a classic case of getting things right for the wrong reason. GTAP predicts a large increase in the quantity of beef exports because of a complete removal of the tariff quota in the simulation. In actuality, the trade agreement between Australia and the United States left the quota largely intact, and the increase in trade value was due to a worldwide increase in the price of beef. If we instead ran GTAP without simulating a complete elimination of all trade barriers (a complete removal is a reasonable approximation for the other industries and trade agreements we evaluate), the GTAP model would be unable to capture the increased value of exports of cattle meat. Because it is questionable whether any trade model should be expected to capture such an increase in the worldwide price, we exclude the industry from our main analysis in the paper.

Although our robustness appendix shows that our overall results are largely unchanged by outliers, base years, weighting schemes, and how we define growth, it also shows that for some individual country pairs, these choices substantially alter the apparent accuracy of various predictions and methodologies. Thus, researchers and policy makers need to think carefully about how the performance of their models should be evaluated. Beyond that, we need to actually carry out such evaluations, so that shortcomings can be identified and our models can continue to improve.