NBER WORKING PAPER SERIES

# RICARDO'S THEORY OF COMPARATIVE ADVANTAGE: OLD IDEA, NEW EVIDENCE

Arnaud Costinot
Dave Donaldson

Working Paper 17969
http://www.nber.org/papers/w17969

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2012

We thank Pol Antràs, Chang-Tai Hsieh, and Esteban Rossi-Hansberg for comments and Meredith McPhail and Cory Smith for excellent research assistance. This paper has been prepared for the 2012 American Economic Review Papers and Proceedings. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Ricardo's Theory of Comparative Advantage: Old Idea, New Evidence
Arnaud Costinot and Dave Donaldson
NBER Working Paper No. 17969
April 2012
JEL No. F11,Q11,Q15,Q17,R14

**ABSTRACT**

When asked to name one proposition in the social sciences that is both true and non-trivial, Paul Samuelson famously replied: 'Ricardo's theory of comparative advantage'. Truth, however, in Samuelson's reply refers to the fact that Ricardo's theory of comparative advantage is mathematically correct, not that it is empirically valid. The goal of this paper is to assess the empirical performance of Ricardo's ideas. We use novel agricultural data that describe the productivity in 17 crops of 1.6 million parcels of land in 55 countries around the world. Crucially, this dataset contains information about the productivity of each parcel of land in all crops, not just those that are currently being grown. This direct information about relative productivity differences across economic activities allows us to compute, for the first time, the output predicted by Ricardo's theory of comparative advantage. Despite all of the real-world considerations from which this theory abstracts, we find that Ricardo's theory of comparative advantage has significant explanatory power in the data, at least within the scope of our analysis.

Arnaud Costinot
Department of Economics
MIT, E52-243B
50 Memorial Drive
Cambridge MA 02142-1347
and NBER
costinot@mit.edu

Dave Donaldson
MIT Department of Economics
50 Memorial Drive, E52-243G
Cambridge, MA 02142-1347
and NBER
ddonald@mit.edu

## 1 Introduction

The anecdote is famous. A mathematician, Stan Ulam, once challenged Paul Samuelson to name one proposition in the social sciences that is both true and non-trivial. His reply was: 'Ricardo's theory of comparative advantage'; see Paul Samuelson (1995, p. 22). Truth, however, in Samuelson's reply refers to the fact that Ricardo's theory of comparative advantage is mathematically correct, not that it is empirically valid. The goal of this paper is to assess the empirical performance of Ricardo's ideas.

To bring Ricardo's ideas to the data, one must overcome a key empirical challenge. Suppose, as Ricardo's theory of comparative advantage predicts, that different factors of production specialize in different economic activities based on their relative productivity differences. Then, following Ricardo's famous example, if English workers are relatively better at producing cloth than wine compared to Portuguese workers, England will produce cloth, Portugal will produce wine, and at least one of these two countries will be completely specialized in one of these two sectors. Accordingly, the key explanatory variable in Ricardo's theory, relative productivity, cannot be directly observed.

This identification problem is emphasized by Alan Deardorff (1984) in his review of empirical work on the Ricardian model of trade (p. 476): "Problems arise, however, most having to do with the observability of [productivity by industry and country]. The...problem is implicit in the Ricardian model itself...[because] the model implies complete specialization in equilibrium... This in turn means that the differences in labor requirements cannot be observed, since imported goods will almost never be produced in the importing country." A similar identification problem arises in the labor literature in which the self-selection of individuals based on comparative advantage is often referred to as the Roy model. As James Heckman and Bo Honore (1990) have shown, if general distributions of worker skills are allowed, the Roy model—and hence Ricardo's theory of comparative advantage—has no empirical content. Econometrically speaking, the Ricardian model is not nonparametrically identified.

How can one solve this identification problem? One possibility consists in making untestable functional form assumptions about the distribution of productivity across different factors of productions and economic activities. These assumptions can then be used to relate productivity levels that are observable to those that are not. In a labor context, a common strategy is to assume that workers' skills are log-normally distributed. In a trade context, building on the work of Jonathan Eaton and Samuel Kortum (2002), Arnaud Costinot, Dave Donaldson, and Ivana Komunjer (2011) have shown how the predictions of the Ricardian model can be tested by assuming that productivity levels are independently drawn from Fréchet distributions across countries and industries.

This paper proposes an alternative empirical strategy that does not rely on identification by functional form. Our basic idea, as in Arnaud Costinot and Dave Donaldson (2011), is to focus on agriculture, a sector of the economy in which scientific knowledge of how essential inputs such as water, soil and climatic conditions map into outputs is uniquely well understood. As a consequence of this knowledge, agronomists are able to predict how productive a given parcel of land, which will we refer to as a 'field', would be were it to be used to grow any one of a set of crops. In this particular context, the econometrician therefore knows the productivity of a field in *all* economic activities, not just those in which it is currently employed.

Our strategy can be described as follows. We first establish how, according to Ricardo's theory of comparative advantage, total output of various crops should vary across countries as a function of: ($i$) the vector of productivity of the fields that countries are endowed with and ($ii$) the producer prices that determine the allocation of fields across crops.[1] We then combine these theoretical predictions with productivity and price data from the Food and Agriculture Organization's (FAO). Our dataset consists of 17 major agricultural crops and 55 major agricultural countries. Using this information, we can compute predicted output levels for all crops and countries in our sample and ask: How do predicted output levels compare with those that are observed in the data?

Our empirical results show that the output levels predicted by Ricardo's theory of comparative advantage agree reasonably well with actual data on worldwide agricultural production. Despite all of the real-world considerations from which Ricardo's theory abstracts, a regression of log output on log predicted output has a (precisely estimated) slope of 0.21. This result is robust to a series of alternative samples and specifications.

The rest of the paper is organized as follows. Section I derives predicted output levels in an economy where factor allocation is determined by Ricardian comparative advantage. Section II describes the data that we use to construct measures of both predicted and actual output. Section III compares predicted and observed output levels and Section IV offers some concluding remarks.

## 2 Ricardian Predictions

The basic environment is the same as in Costinot (2009). We consider a world economy comprising $c = 1, ..., C$ countries, $g = 1, ..., G$ goods, and $f = 1, ..., F$ factors of production. In our empirical analysis, a good will be a crop and a factor of production will be a parcel of land or 'field'. Factors of production are immobile across countries and perfectly mobile across sectors. $L_{cf} \geq 0$ denotes the inelastic supply of factor $f$ in country $c$. Factors of production are perfect substitutes within each country and sector, but vary in their productivity $A_{cf}^{g} \geq 0$. Total output of good $g$ in country $c$ is given by

$$Q_c^g = \sum_{f=1}^{F} A_{cf}^g L_{cf}^g,$$

where $L_{cf}^g$ is the quantity of factor $f$ allocated to good $g$ in country $c$. The variation in $A_{cf}^g$ is the source of Ricardian comparative advantage. If two factors $f_1$ and $f_2$ located in country $c$ are such that $A_{cf_2}^{g_2}/A_{cf_2}^{g_1} > A_{cf_1}^{g_2}/A_{cf_1}^{g_1}$ for two goods $g_1$ and $g_2$, then field $f_2$ has a comparative advantage in good $g_2$.[2]

[1] In line with Ricardo's theory of comparative advantage, the focus of our paper is on the supply-side of the economy, not the demand-side considerations that would ultimately pin down prices around the world.

[2] The present model, like the Roy model in the labor literature, features multiple factors of production. In international trade textbooks, by contrast, Ricardo's theory of comparative advantage is associated with models that feature only one factor of production, labor. In our view, this particular formalization of Ricardo's ideas is too narrow

Throughout this paper, we focus on the supply-side of this economy by taking producer prices $p_c^g \geq 0$ as given. We assume that the allocation of factors of production to each sector in each country is efficient and solves

$$\max_{L_{cf}^g} \left\{ \sum_{c=1}^{C} \sum_{g=1}^{G} p_c^g Q_c^g \middle| \sum_{g=1}^{G} L_{cf}^g \leq L_{cf} \right\}.$$

Since there are constant returns to scale, a competitive equilibrium with a large number of profit-maximizing firms would lead to an efficient allocation. Because of the linearity of aggregate output, the solution of the previous maximization problem is easy to characterize. As in a simple Ricardian model of trade with two goods and two countries, each factor should be employed in the sector that maximizes $A_{cf}^g p_c^g$, independently of where other factors are being employed.

Assuming that the efficient allocation is unique,[3] we can express total output of good $g$ in country $c$ at the efficient allocation as

$$Q_c^g = \sum_{f \in \mathcal{F}_c^g} A_{cf}^g L_{cf}, \tag{1}$$

where $\mathcal{F}_c^g$ is the set of factors allocated to good $g$ in country $c$:

$$\mathcal{F}_c^g = \left\{ f = 1, ...F | \frac{A_{cf}^g}{A_{cf}^{g'}} > \frac{p_c^{g'}}{p_c^g} \text{ if } g\text{'} \neq g \right\}. \tag{2}$$

Equations (1) and (2) capture Ricardo's idea that relative rather than absolute productivity differences determines factor allocation, and in turn, the pattern of international specialization.

## 3 Data

To assess the empirical performance of Ricardo's ideas we need data on actual output levels, which we denote by $\widetilde{Q}_c^g$, as well as data to compute predicted output levels, which we denote by $Q_c^g$ in line with Section I. According to equations (1) and (2), $Q_c^g$ can be computed using data on productivity, $A_{cf}^g$, for all factors of production $f$; endowments of different factors, $L_{cf}$; and producer prices, $p_c^g$. We describe our construction of such measures here. Since the predictions of Ricardo's theory of comparative advantage are fundamentally cross-sectional in nature, we work with the data from 1989 only; this is the year in which the greatest overlap in the required measures is available.

We use data on both agricultural output ($\widetilde{Q}_c^g$) and producer prices ($p_c^g$) by country and crop from FAOSTAT. Output is equal to quantity harvested and is reported in tonnes. Producer prices are

for empirical purposes. The core message of Ricardo's theory of comparative advantage is not that labor is the only factor of production in the world, but rather that relative productivity differences, and not absolute productivity differences, are the key determinant of factor allocation. As argued below, the present model captures exactly that idea.

[3] In our empirical analysis, 2 out of the 101,757 grid cells in Brazil—the empirical counterparts of factors $f$ in the model—are such that the value of their marginal products $A_{cf}^g p_c^g$ is maximized in more than one crop. Thus the efficient allocation is only unique up to the allocation of these two Brazilian grid cells. Dropping these two grid cells has no effect on the coefficient estimates presented in Table 1.

equal to prices received by farmers net of taxes and subsidies and are reported in local currency units per tonne. Imperfect data reporting to the FAO means that some output and price observations are missing. We first work with a sample of 17 crops and 55 countries that is designed to minimize the number of missing observations.[4] In the remaining sample, whenever output data is missing we assume that there is no production of that crop in that country. Similarly, whenever price data is unreported for a given observation, both quantity produced and area harvested are also reported as zero in the FAO data. In these instances, we therefore replace the missing price entry with a zero.[5]

Our data on productivity ($A^g_{cf}$) come from version 3.0 of the Global Agro-Ecological Zones (GAEZ) project run by IIASA and the FAO (IIASA/FAO, 2012). We describe this data in detail in Costinot and Donaldson (2011) but provide a brief description here; see also Nathan Nunn and Nancy Qian (2009). The GAEZ project aims to make agronomic predictions about the yield that would obtain for a given crop at a given location for all of the world's major crops and all locations on Earth. Data on natural inputs (such as soil characteristics, water availability, topography and climate) for each location are fed into an agronomic model of crop production with distinct parameters for each variety of each crop. These models condition on a level of variable inputs and GAEZ makes available the output from various scenarios in which different levels of variable inputs are applied. We use the scenario that corresponds to a 'mixed' level of inputs, where the farmer is assumed to be able to apply inputs differentially across sub-plots within his or her location, and in which irrigation is available. It is important to stress that the thousands of parameters that enter the GAEZ model are estimated from countless field and lab experiments, not from statistical relationships between observed country-level output data (such as that from FAOSTAT which we use here to construct $\widetilde{Q}^g_c$) and natural inputs.

The spatial resolution of the GAEZ data is governed by the resolution of the natural input whose resolution is most coarse, the climate data. As a result the GAEZ productivity predictions are available for each 5 arc-minute grid cell on Earth. The land area of such a cell varies by latitude but is 9.2 by 8.5 km at the Tropics. The median country in our dataset contains 4,817 grid cells but a large country such as the U.S. comprises 157,797 cells. Since the grid cell is the finest unit of spatial heterogeneity in our dataset we take each grid cell to be a distinct factor of production $f$ and the land area of each grid cell to be the associated endowment, $L_{cf}$. Hence our measure of the productivity of factor $f$ if it were to produce crop $g$ in country $c$, $A^g_{cf}$, corresponds to the GAEZ project's predicted 'total production capacity (tones/ha)'. We match countries (at their

[4] The countries are: Argentina, Australia, Austria, Bangladesh, Bolivia, Brazil, Bulgaria, Burkina Faso, Cambodia, Canada, China, Colombia, Democratic Republic of the Congo, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Finland, France, Ghana, Honduras, Hungary, Iceland, Indonesia, Iran, Ireland, Israel, Jamaica, Kenya, Laos, Lebanon, Malawi, Mozambique, Namibia, Netherlands, Nicaragua, Norway, Paraguay, Peru, Poland, Romania, South Africa, Spain, Suriname, Sweden, Togo, Trinidad and Tobago, Tunisia, Turkey, USSR, United States, Venezuela, Yugoslavia and Zimbabwe. The crops are: barley, cabbages, carrots and turnips, cassava, coconuts, seed cotton, groundnuts (with shell), maize, onions (dry), rice (paddy), sorghum, soybeans, sugar cane, sweet potatoes, tomoatoes, wheat, potatoes (white).

[5] We have also experimented with replacing missing prices by their world averages across producing countries adjusted for currency differences. The empirical results in Table 1 are insensitive to this alternative.

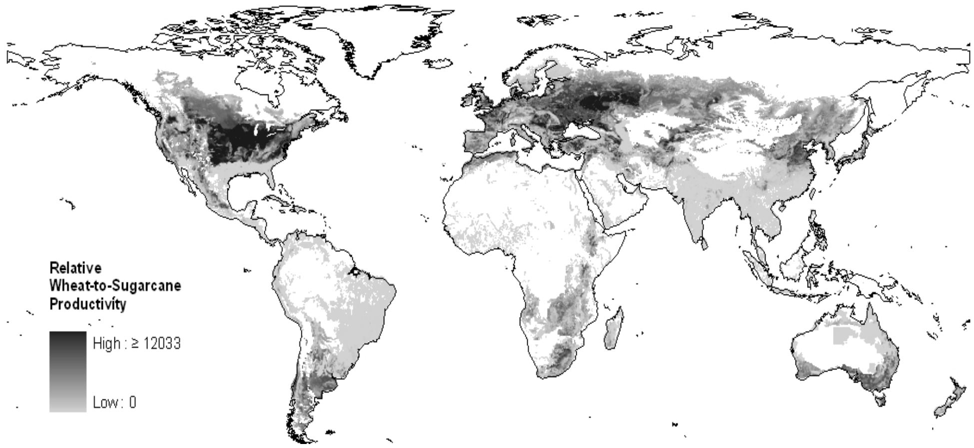


**Figure 1: An Example of Relative Productivity Differences.** Notes: Ratio of productivity in wheat (in tonnes/ha) relative to productivity in sugarcane (in tonnes/ha). Areas shaded white have either zero productivity in wheat, or zero productivity in both wheat and sugarcane. Areas shaded dark, with the highest value (">12,033"), have zero productivity in sugarcane and strictly positive productivity in wheat. Source: GAEZ project.

1989 borders) to grid cells using GIS files on country borders from the Global Administrative Areas database.

A sample of the GAEZ predictions can be seen in Figure 1. Here we plot, for each grid cell on Earth, the predicted relative productivity in wheat compared to sugarcane (the two most important crops by weight in our sample). As can be seen, there exists a great deal of heterogeneity in relative productivity throughout the world, even among just two of our 17 crops. In the next section we explore the implications of this heterogeneity—heterogeneity that is at the core of Ricardo's theory of comparative advantage—for determining the pattern of international specialization across crops.

# 4 Empirical Results

We are now ready to bring Ricardo's ideas to the data. To overcome the identification problem highlighted by Deardorff (1984) and Heckman and Honore (1990), we take advantage of the GAEZ data, together with the other data described in Section II, to predict the amount of output ($Q_c^g$) that country $c$ should produce in crop $g$ according to Ricardo's theory of comparative advantage, i.e. according to equations (1) and (2). We then compare these predicted output levels to those that are observed in the data ($\widetilde{Q}_c^g$).

In the spirit of the 'slope tests' in the Heckscher-Ohlin-Vanek literature, see Donald Davis and David Weinstein (2001), we implement this comparison by simply regressing, across countries and crops, data on actual output on measures of predicted output. Like Davis and Weinstein (2001), we will assess the empirical performance of Ricardo's ideas by studying whether $(i)$ the slope coefficient in this regression is close to unity and $(ii)$ the coefficient is precisely estimated. Compared to these authors, however, we have little confidence in our model's ability to predict *absolute* levels of output. The reason is simple: the model presented in Section II assumes that the only goods produced (using land) in each country are the 17 crops for which GAEZ productivity data are available. In reality there are many other uses of land, so the aggregate amount of land used to grow the 17 crops in our study is considerably lower than that assumed in our analysis. To circumvent this problem, we simply estimate our regressions in logs.[6] Since the core aspect of Ricardian comparative advantage lies in how *relative* productivity levels predict *relative* quantities, we believe that a comparison of logarithmic slopes captures the essence of what the model described in Section I can hope to predict in this context.

Our empirical results are presented in Table 1. All regressions include a constant and use standard errors that are adjusted for clustering by country to account for potential within-country (across crop) correlation in data reporting and model misspecification. Column (1) contains our baseline regression. The estimated slope coefficient is 0.212 and the standard error is small (0.057).[7] While the slope coefficient falls short of its theoretical value (one), it remains positive and statistically significant.

The fact that Ricardo's theory of comparative advantage does not fit the data perfectly should not be surprising. First, our empirical exercise focuses on land productivity and abstracts from all other determinants of comparative costs (such as factor prices that differ across countries and factor intensities that differ across crops) that are likely to drive agricultural specialization throughout the world. Second, the fit of our regressions does not only depend on the ability of Ricardo's theory to predict relative output levels conditional on relative productivity levels, but also on the ability of agronomists at the GAEZ project to predict productivity levels in each of 17 crops at 5 arc-minute grid cells throughout the world conditional on the (counterfactual) assumption that all countries share a common agricultural technology.[8] Third, while the spatial resolution of the GAEZ predictions is considerably finer than the typical approach to cross-country data in the trade literature (in which countries are homogeneous points), 5 arc-minute grid cells are still very coarse in an absolute sense. This means that there is likely to be a great deal of potential within-country heterogeneity that is being smoothed over by the GAEZ agronomic modeling. Yet despite these limitations of our analysis, Ricardo's theory of comparative advantage still has significant explanatory power in the data, as column (1) illustrates.

[6] In order to measure the gains from the economic integration of U.S. agricultural markets between 1880 and 2000, Costinot and Donaldson (2011) have developed a methodology that uses additional data on aggregate land use to correct for this problem. Applying that correction is computationally challenging here, due to the large number of fields in most countries, and is beyond the scope of the present paper.

[7] In our logarithmic specification all observations in which either output or predicted output are zero must be omitted. Out of the total of 935 potential observations (55 countries and 17 crops), 296 have zero output and 581 have zero predicted output—that is our Ricardian model predicts more complete specialization that there is in the data. This should not be surprising given the potential for more spatial heterogeneity to exist in agricultural reality than can be modeled (due to data limitations) by GAEZ. In all, 349 observations have both non-zero output and non-zero predicted output and are hence included in the regression in column (1). We have explored a number of potential adjustments to correct the results in column (1) for these missing observations, including a Tobit regression (where the coefficient is 0.213 and the s.e. is 0.057) and adding one to all observations prior to taking logs (coefficient 0.440; s.e. 0.031).

[8] The methodology developed in Costinot and Donaldson (2011) uses data on harvested area to allow for and estimate unrestricted crop-and-region productivity shocks. Again, because of the high number of fields per country applying this correction to the current paper is computationally challenging.

**Table 1: Comparison of Actual Output to Predicted Output**

| Dependent variable: | log (output) | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| log (predicted output) | 0.212*** | 0.244*** | 0.096** | 0.143** | 0.273*** |
| | (0.057) | (0.074) | (0.038) | (0.062) | (0.074) |
| sample | all | all | all | major countries | major crops |
| fixed effects | none | crop | country | none | none |
| observations | 349 | 349 | 349 | 226 | 209 |
| R-squared | 0.06 | 0.26 | 0.54 | 0.04 | 0.07 |

Notes: All regressions include a constant. Standard errors clustered by country are in parentheses. ** indicates statistically significant at 5% level and *** at the 1% level.

Columns (2) and (3) explore the robustness of our baseline estimate in column (1) to the inclusion of crop and country fixed effects, respectively. The rationale for these alternative specifications is that there may be crop- or country-specific tendencies for misreporting or model error. Such errors may be economic in nature if, say, some countries had higher intra-national price distortions, or agronomic in nature if, say, the GAEZ model predictions were relatively more accurate for some crops than others. Including such fixed effects can reduce the slope coefficient (to as low as 0.096, in column (3)) but these estimates are still statistically significantly different from zero. Thus the results in columns (2) and (3) show that Ricardo's theory of comparative advantage continues to have explanatory power whether focusing on the across-country variation, as in column (2), or the across-crop variation, as in column (3).

Finally, columns (4) and (5) investigate the extent to which our estimates are driven by particular components of the sample. Column (4) estimates the slope only among the 28 countries that are at or above the median in terms of agricultural production (by weight). And column (5) estimates the slope only on the 9 crops that are the most important (by weight) in global production. In both cases the estimated slope coefficient is similar (within one standard error) to our baseline estimate in column (1).

## 5 Concluding Remarks

Ricardo's theory of comparative advantage is one of the oldest and most distinguished theories in economics. But it is a difficult theory to bring to the the data. To do so using conventional data sources, one needs to make untestable functional form assumptions about how productive a given factor of production would be at the activities it is currently, and deliberately, not doing. In this paper we have argued that the predictions of agronomists—i.e., the scientists who specialize in modeling how agricultural crops would fare under a wide range of possible growing conditions—can be used to provide the missing data that make Ricardo's ideas untestable in conventional settings.

We have combined the data from a particular group of agronomists, those working on the GAEZ project as part of the FAO, along with producer price data from the FAO, to assess the empirical performance of Ricardo's ideas across 17 agricultural crops and 55 major agriculture-producing countries in 1989. We have asked a simple question: How do output levels predicted by Ricardo's theory compare to those that are observed in the data? Despite all of the real-world considerations from which Ricardo's theory abstracts, we find that a regression of log output on log predicted output has a (precisely estimated) slope of 0.21. Ricardo's theory of comparative advantage is not just mathematically correct and non-trivial; it also has significant explanatory power in the data, at least within the scope of our analysis.

## REFERENCES


**Costinot, Arnaud. 2009.** "An Elementary Theory of Comparative Advantage." *Econometrica* 77(4): 1165-1192.

**Costinot, Arnaud, and Dave Donaldson.** 2011. "How Large Are the Gains from Economic Integration? Evidence from US Agriculture 1880-2000." Unpublished.

**Costinot, Arnaud, Dave Donaldson, and Ivana Komunjer.** 2011. "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas." *Review of Economic Studies* forthcoming.

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

**Deardorff, Alan.** 1984. "Testing Trade Theories and Predicting Trade Flows." Handbook of International Economics, Volume 1, R.W. Kenen and P.B. Jones editors. North Holland, Amsterdam.

**Eaton, Jonathan, and Samuel Kortum**. 2002. "Technology, Geography, and Trade." *Econometrica* 70(5): 1741-1779.

**Heckman, James, and Bo Honore.** 1990. "The Empirical Content of the Roy Model." *Econometrica* 58(5): 1121-1149.

**IIASA/FAO**. 2012. *Global Agro-ecological Zones (GAEZ v3.0).* IIASA, Laxenburg, Austria and FAO, Rome, Italy.

**Nunn, Nathan and Nancy Qian.** 2011. "The Potato's Contribution to Population and Urbanization: Evidence from a Historical Experiment," *Quarterly Journal of Economics*,126(2), pp. 593-650.

**Samuelson, Paul A**. 1995. "The Past and the Future of International Trade Theory." In *New Directions in Trade Theory*, edited by Jim Levinsohn, Alan V. Deardorff, and Robert M. Stern, pp 17-22. Ann Arbor: The University of Michigan Press.