Trade data defects and their implications for statistical inference

Lukas Linsi
University of Groningen

Brian Burgoon
University of Amsterdam

Daniel Mügge
University of Amsterdam

Draft version to be presented at the virtual 2020 IPES Conference, November 13-14. Please do not cite or circulate without the permission of the authors.

Abstract. Trade statistics are widely used in studies and policymaking focused on economic interdependence. Yet researchers in International Relations (IR) and International Political Economy (IPE) have largely disregarded a fundamental defect of this data: the *mirror problem* of trade statistics. Bilateral trade flows are usually recorded twice: by the sending economy as an export and by the receiving one as an import. In theory, these two values should match. In practice, discrepancies between them are large and pervasive. Most studies circumvent the mirror problem by using only one entry (usually the importer-based figures) while disregarding the other. This is problematic. It is not self-evident that one measure is consistently more accurate than the other. By doing so, data users thus give trade figures a misleading veneer of accuracy. Against this background, this article makes three contributions: first, we quantify the mirror problem in both dyadic and monadic settings. Second, we investigate the origins of the mirror problem, leveraging statistical analyses as well as archival records and interviews with statistical experts. Third, we illustrate the possible implications of the mirror problem for statistical inference through replications covering diverse topics in IR and IPE. We find that accounting for the mirror problem can variably strengthen, undermine or overturn the conclusions of such analyses. The findings underscore the severity of measurement problems in IR and IPE, and we suggest particular ways to engage with them.

Keywords. Trade statistics; measurement error; mirror statistics; economic interdependence

Word count: 13,939 (incl. references)

1. Introduction

Trade statistics are prominent in global economic governance and international relations research. Cross-border trade remains the bedrock of economic ties between nation-states, and measures of it inform trade policies and development strategies throughout the world. Among international relations (IR) and international political economy (IPE) scholars, import and export figures are the most commonly used measures of economic interdependence, and are crucial to understanding the character, origins, and implications of economic globalization. Such trade figures also feed causal analyses, for example the study of domestic and international political struggles or institutional developments in the global political economy.

Research designs and statistical modelling employed to study the origins and consequences of trade have become ever more advanced, and extensively debated. While such debate focuses mainly on issues of causal identification, scholars and policymakers have almost entirely disregarded major defects of trade data itself. The International Monetary Fund (IMF) or the Organization for Economic Cooperation and Development (OECD), which publish international economic statistics, acknowledge that trade data quality can be wanting.[1] Digital trade or grey-market economic transactions are notoriously hard to capture, for example.[2] At times, statistical offices in different countries use disparate valuation methods or disagree about the ultimate origin or destination of merchandise.[3]

Such measurement uncertainties surface in so-called "mirror statistics." Trade flows are in principle recorded twice, once as an export by the sending economy, and once as an import by the

[1] International Monetary Fund 1987; UNECE, Eurostat, and OECD 2011.
[2] Gaspareniene, Remeikiene, and Schneider 2015.
[3] Markhonko 2014.

receiving one. The IMF's Direction of Trade Statistics (DOTS) database[4] is the most widely used resource for bilateral trade statistics in IR and IPE research, and it provides both figures. If they were very similar—as they should be—mirror statistics would not raise significant questions. Yet discrepancies in mirror statistics are large and persistent, even between countries with highly developed statistical systems.

This mirror problem, as we call it, reveals the substantial uncertainty in trade statistics, which poses a potential challenge to anyone exploring the character, origins, or implications of trade.[5] Trade statisticians and economists have long recognized this issue,[6] and they have proposed various statistical remedies, such as the estimation of mirror averages weighted by inferred reporter reliability in the BACI[7] or OECD BIMTS[8] data sets. While we recognize these efforts, we show that they cannot tackle the mirror or other data problems fully and remain too limited in their coverage for many analytical purposes in IR and IPE. Meanwhile, notwithstanding awareness among statisticians and economists, most IR and IPE work ignores data defects altogether.[9] Using the most widely available trade statistics, based on import values alone, most IR/IPE scholars implicitly trust those measures as better than export figures. This assumption, as we argue below, frequently does not hold.

The mirror problem and the lack of attention to it in IR and IPE research provide probable cause for this article's mission: to understand better how measurement problems in trade data affect the validity and reliability of IR and IPE research focused on trade. We pursue this mission

[4] Available here: http://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85 [last accessed: 1 August 2020]

[5] Cf. Morgenstern 1963; International Monetary Fund 1987; Schultz 2015; Linsi and Mügge 2019.

[6] Ely 1961; Morgenstern 1963; Bhagwati 1964; Bhagwati 1967; Yeats 1978; Yeats 1990; Gaulier and Zignago 2010.

[7] Available here: http://www.cepii.fr/cepii/en/bdd_modele/presentation.asp?id=37 [last accessed: 5 August 2020]

[8] Available here: https://stats.oecd.org/Index.aspx?DataSetCode=BIMTS_CPA [last accessed: 5 August 2020]

[9] Studies that do discuss data problems are the exceptions to the rule, for example Barbieri, Keshk, and Pollins 2009; Gleditsch 2010; Boehmer, Jungblut, and Stoll 2011; Schultz 2015.

in three steps. First, we construct measures that quantify the mirror problem in both dyadic terms (between pairs of states) and monadic terms (concerning a country's aggregate trade). This yields two publicly available datasets of errors in common trade measures.[10] These datasets reveal large and persistent discrepancies that are not confined to specific countries or regions of the world.

Second, we explore the sources of these measurement problems. Archival and interview research with leading trade statisticians highlights how complex data collection is for trade measurement. Quantitative analysis of mirror discrepancies reveals their sources to be many and uncertain: we find systematic biases, but a substantial portion of discrepancies remains unexplained even in the most comprehensive fixed-effects models. We cannot, therefore, simply model mirror discrepancies out of our data.

Third, we explore the implications of the mirror problem for IR and IPE research. We replicate five studies chosen to cover a wide variety of topics and statistical setups.[11] They include the effects of economic globalization on welfare states and those of multilateral institutions on actual trade interdependence. We also consider international diplomacy and security issues, such as the link between trade, geopolitical alignments and military conflicts. The studies feature trade as both explanation and outcome, in both security and political economy issues, and in both dyadic and monadic settings.

Our replications reveal that measurement uncertainty is a consequential issue. Accounting for measurement error can significantly strengthen or altogether wash-out the statistical significance of previous results. It frequently yields substantial changes in the magnitude of estimated effects;

[10] The version accompanying this article covers the years 1950-2014. We intend to update the data set periodically.
[11] These studies are those by Kastner (2014), Rose (2004), Goldstein, Rivers and Tomz (2007), Barbieri and Reuveny (2005), and Garrett and Mitchell (2001).

in some cases, it reverses their direction. Data transformation and aggregation can attenuate the problem, but not altogether erase it.

2. The use of trade statistics in IR and IPE research

Cross-border commerce stands central in international economic relations. To gauge how widely IR and IPE scholars use statistics about it, we reviewed all articles published between 2013 and 2017 in leading IR and political science journals: International Organization, International Studies Quarterly, World Politics, Journal of Conflict Resolution, Journal of Politics, and European Journal of Political Research. In total, 108 articles used trade data (slightly more than 1 out of 15), almost always at the country-level. Trade flows appear in four primary modes of analysis: monadic-country (e.g. total imports/exports of a country); monadic-product (e.g. imports/exports of goods in specific product categories); dyadic-country (e.g. total flows among country-pairs); and dyadic-product (e.g. bilateral flows in specific product categories). Of the 108 studies, 49 used dyadic-country data and 46 monadic-country data. Product-level trade data remains rare in IR and political science.[12] We identified only eleven studies using monadic-product and five employing dyadic-product data (see Figure 1).[13]

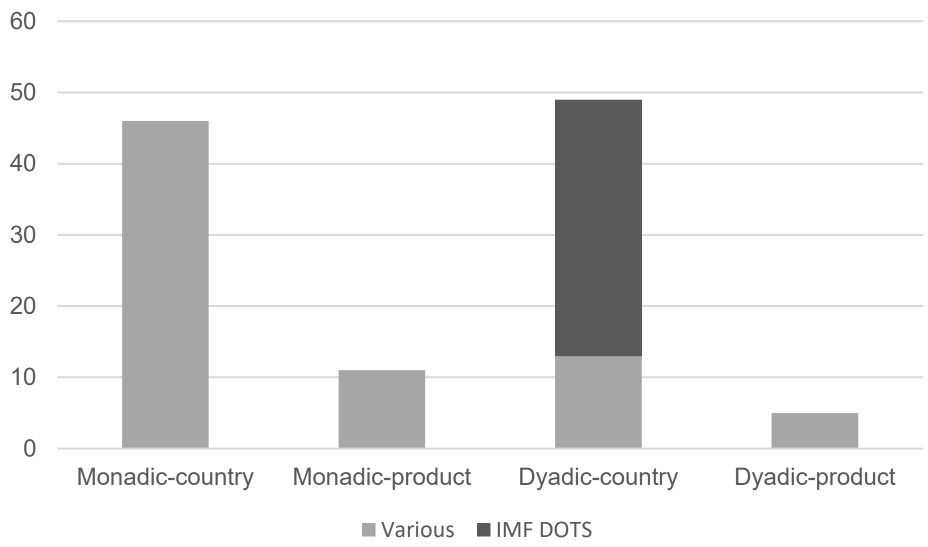
Well-known data-gathering bodies dominate as data sources. More than 60 percent of the monadic-country studies rely on the World Bank's (WB) World Development Indicators database, with the OECD and WB national accounts data as ultimate sources. Eleven percent draw on Penn World Tables (mostly UN sources); and remainder comes from US government and other sources (e.g. Eurostat), or is not specified (15 percent). Monadic-product level studies draw primarily on

[12] Kim, Liao, and Imai 2020.

[13] Some articles use more than one type of trade data. Our categorization excludes one study that uses firm-level trade data.

the World Integrated Trade Solution (WITS) database (63 percent), with only one study using the reconciled figures from the BACI dataset,[14] more commonly used in economics and discussed in more detail below. Two of the five dyadic-product analyses also use WITS and the other specialized national sources. Among dyadic-country level analyses, almost three quarters rely on IMF DOTS (either directly, or by using the Gleditsch or Correlates of War (COW) databases, both of which build on IMF DOTS). The remainder comes from UN, NBER and sundry national (e.g. US government) or regional sources (e.g. Eurostat). These studies tend to take trade-data quality for granted; critical discussion or analysis of it is sparse beyond the occasional general disclaimer.

Figure 1. Trade data use in six leading journals, 2013-17



SOURCE: Data collected by authors from journal homepages (details in text).

The subfield in which trade measurement has been discussed most extensively focuses on the link between trade and violent conflict. Scholars there have not addressed the mirror problem directly but raised a closely related one: missing data. COW and Gleditsch's expanded bilateral

[14] Gaulier and Zignago 2010.

trade dataset (both based on IMF DOTS) are prominent in trade and conflict research. COW treats missing trade values as missing; the Gleditsch dataset assumes that missing data reflects very little or no trade, justifying setting these values to zero. Disagreements among the compilers of the two datasets have generated important debates,[15] not least because the treatment of missing values fundamentally affects the statistical results.[16]

Measurement uncertainty in trade data, however, is not limited to unobserved values. Some economists recognized such more general data defects long ago. In *On the Accuracy of Economic Observations*, Oskar Morgenstern noted in 1950 that "[writers] on all phases of foreign trade will have to assume the burden of proof that the figures on commodity movements are good enough to warrant the manipulation and the reasoning to which they are customarily subject."[17] More recently, Bhagwati analyzed how deliberate over- or under-invoicing of trade biased balance of payments (BOP) data in the 1960s and 1970s.[18] Other scholars lamented that discrepancies in bilateral trade records were "often considerable."[19] Analyzing African trade statistics, Yeats claimed that "these data cannot be relied on to indicate the level, composition, or even direction and trends in … trade."[20] Studies from other regions raised similar concerns.[21] Statistical agencies and international organizations, too, have highlighted the mirror problem for some time,[22] even if it remains unresolved.[23]

[15] Barbieri, Keshk, and Pollins 2009; Gleditsch 2010; Barbieri and Keshk 2011.
[16] Boehmer, Jungblut, and Stoll 2011.
[17] Morgenstern 1963 [1950], 180.
[18] Bhagwati 1964; Bhagwati 1967.
[19] Yeats 1978, 354; also Ely 1961.
[20] Yeats 1990, 135.
[21] Naya and Morgan 1969; Braml and Felbermayr 2019.
[22] International Monetary Fund 1993; Javorsek 2016; Garber, Peck, and Howell 2018; Office for National Statistics 2020; International Monetary Fund 1987.
[23] Schultz 2015; Linsi and Mügge 2019.

Recent years have seen growing interest in the politics behind the production and use of indicators and statistics in international life.[24] Kerner has highlighted the paucity of foreign direct investment (FDI) data;[25] other work has investigated data defects for BOP statistics more generally.[26] Trying to explain skews in WDI economic policy and debt data, Hollyer, Rosendorff and Vreeland have explored the role of IMF programs and countries' regime types for data transparency.[27] Schultz has shown how territorial disputes increase mirror discrepancies among the respective dyads.[28] Also these studies, however, have neither fully appreciated the mirror problem, nor helped us understand its sources or impact on our inferences.

3. Mirror discrepancies and their uncertain origins

Mirror discrepancies arise when two countries record different values for one and the same flow. How substantial are these discrepancies? And what might explain them? As a first empirical impression, Figure 2 visualizes the United States' merchandise trade deficit with Mexico. US figures show it rising sharply from 1995 to 2007 and stabilizing afterwards, until it rises again from 2015 onwards. In Mexican data, the upward trend continues throughout the period. Since 2013, Mexican figures have consistently exceeded American ones by more than 50 percent.

[24] Broome and Quirk 2015; Kelley and Simmons 2019.

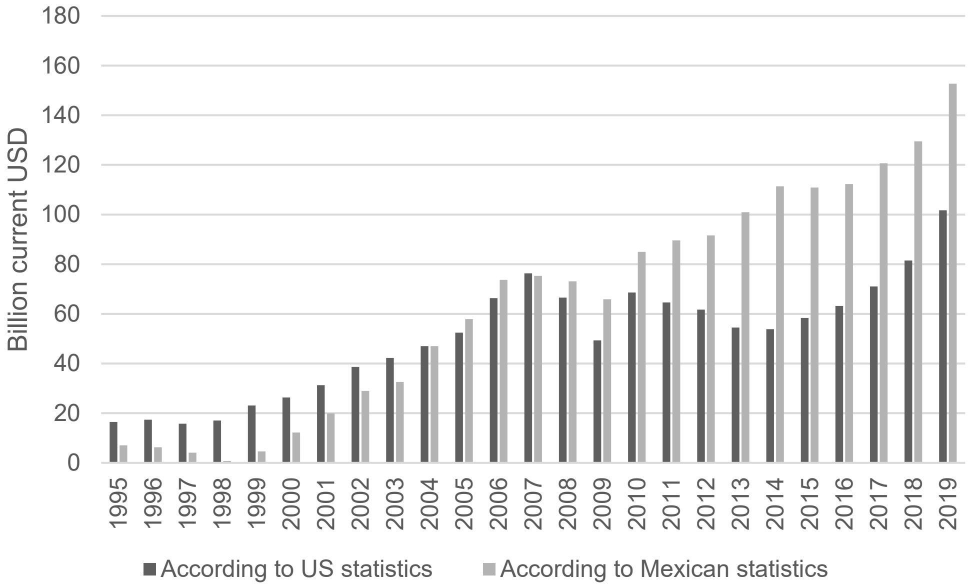
[25] Kerner 2014.

[26] Linsi and Mügge 2019.

[27] Hollyer, Rosendorff, and Vreeland 2011.

[28] Schultz 2015.

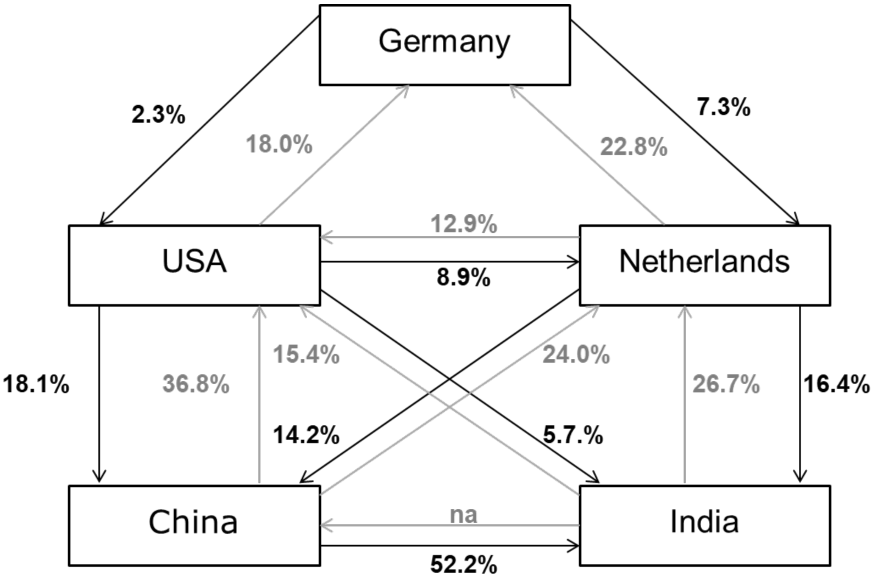
Figure 2. The size of the US merchandise trade deficit with Mexico, 1995-2019



SOURCE: Own calculations based on IMF DOTS database.

Such discrepancies are not limited to the US-Mexico axis, as Figure 3 shows. It depicts trade relations between the USA, Germany, the Netherlands, China and India. The percentages indicate the discrepancy as a share of the total value of recorded imports, averaged over twenty years (1995-2014). For example, on average German and Dutch trade records disagreed about the value of Dutch exports to Germany by more than 20 percent. The message is clear: discrepancies are large and pervasive.

Figure 3. Mirror statistics discrepancies as share of import value, 1995-2014 period-averages



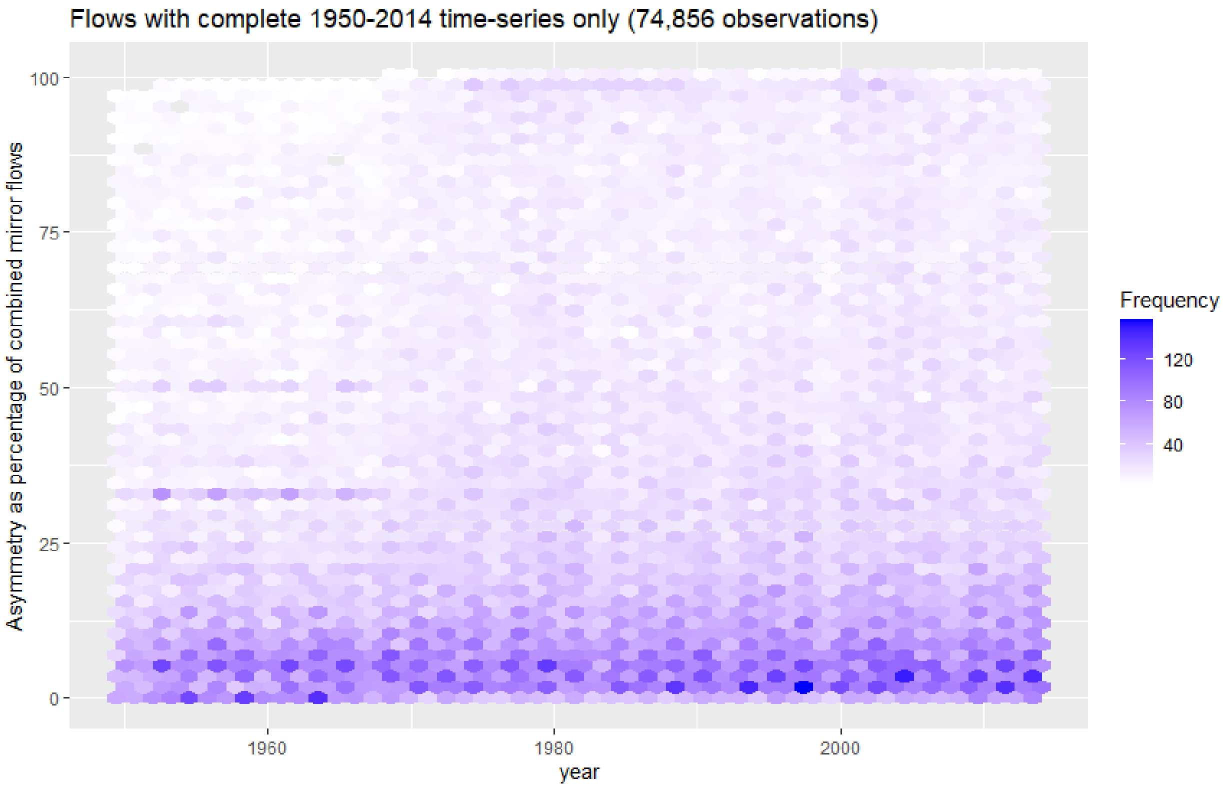
SOURCE: Own calculations based on IMF DOTS.

The density plot in Figure 4 visualizes the discrepancies globally for countries that have consistently reported bilateral trade data since 1950.[29] To standardize, we show them as a share of the two combined estimates of the same flow. The vertical axis gauges their extent for all countries in the sample. The depth of color indicates their frequency: the darker the space, the more country-dyads exhibit a given level of discrepancy in that year. For a significant number of dyads discrepancies are large and have tended to grow rather than decrease with time.[30]

[29] This approach avoids a potential distortion of the overall picture through the addition of newly independent countries, which frequently have a reputation for poor data quality (see Jerven 2013).

[30] Cf. Linsi and Mügge 2019. The mean discrepancy in the sample restricted to dyads with full time series is 29.9 percent (median 19.1), the mean in the full sample 34.5 percent (median 23.6). These are percentages expressed as a share of the combined sum of pairs of mirror flows. In both restricted and full samples, the mean increases gradually over time (from a mean of 24.9 percent in 1950s to 37.1 percent in 2000s in full sample; from 29.9 to 31.5 percent in restricted one). As a complement, appendix figure A1 plots the dyad-specific 1950-2014 period-average ABBA discrepancies in a country-by-country matrix heat plot.

Figure 4. Mirror discrepancy density plot in IMF DOTS

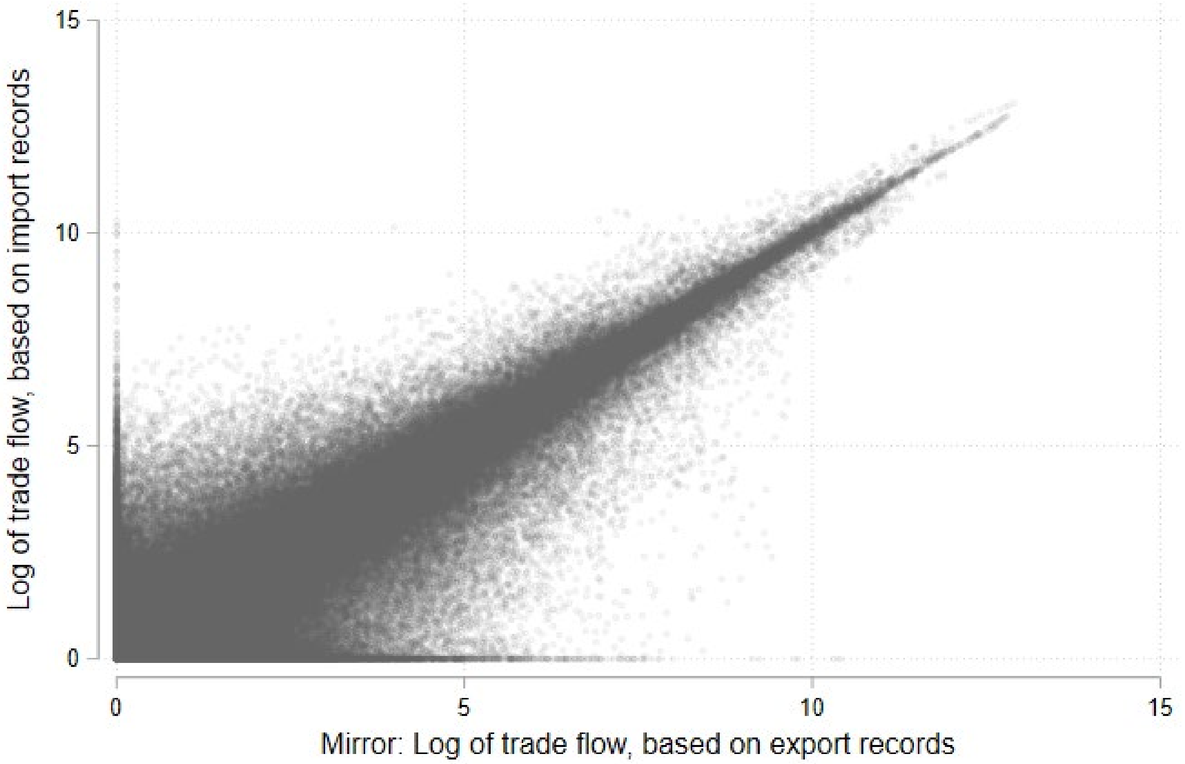


SOURCE: Own analysis based on IMF DOTS database.

Finally, the scatterplot in Figure 5 depicts the discrepancies in gross terms. It plots the log of the import-records based trade value in million current USD's (y-axis) against the mirror entry from export-records (x-axis) for all dyads in the global sample that have two independent mirror records. Also here, deviations are large. Such visualization also shows how the discrepancies are relatively larger at smaller given levels of trade.[31]

[31] The correlation is 0.94 in the overall sample, but only 0.69 in the subsample of flows with an import-record log value below 3 (which accounts for 75 percent of all datapoints), and just 0.34 below a log-value of 1 (representing 58 percent of observations).

Figure 5. Scatterplot of mirror flows in global sample



SOURCE: Own calculations based on IMF DOTS. NOTE: Trade values in million current USD.

3.1. Mirror discrepancies beyond snapshots: ABBA terms

The previous snapshots reveal the scale of mirror discrepancies. To explore their underlying causes and consequences systematically, we need standardized statistical measures of them. The *ABBA terms* we propose to that end, for both dyad-years and country-years, measure differences between what country A reports sending to country B, and what B reports receiving from A (and vice versa). By necessity, we operationalize these ABBA terms differently for dyadic and for monadic data.

The *dyadic ABBA terms* can be defined in a straight-forward way as follows:

$$dyadic\ ABBA_{ab\ t} = |trade_{abA\ t} - trade_{abB\ t}|$$

$$dyadic\ ABBA_{ba\ t} = |trade_{baA\ t} - trade_{baB\ t}|$$

Here “a” and “b” denote the origin and destination of an annual bilateral trade flow, “A” and “B” the countries estimating it, and “t” the year. Per dyad and year, we therefore have two ABBA terms, one for each direction of trade.[32] This initial definition is deliberately simple so that they can be used flexibly and adapted to the analytical context, for example normalizing them by a common denominator. In most of our dyadic analyses below, we use the log value of the size of the inflation-adjusted ABBA discrepancy divided by the sum of the two mirror flows.[33]

Monadic ABBA terms are more complicated, because we have to aggregate the dyadic information. Here we measure the difference between (i) the sum of the value of all import [export] flows recorded by the reporting “home” economy and (ii) the sum of the value of all mirror flows recorded by *partner* countries. We limit ourselves to those observations for which we have two independently recorded estimates, excluding those dyad-years for which one data point is missing or has been merely imputed based on partner records.[34] The basic *monadic ABBA term* can be defined as follows:

[32] Our own analyses work with a unidirectional dyadic database, in which all flows are transformed to $trade_{ba}$ (A’s imports from B) and each country in a dyad-year is entered once as ‘A’ (“home”/receiving economy) and once as ‘B’ (“sending” economy). The public dyadic datasets accompanying this article provide the ABBA information in both unidirectional and bidirectional format.

[33] Formally, log ($\frac{|trade_{baA\ t\ 1967USD} - trade_{baB\ t\ 1967USD)}|}{trade_{baA\ t\ 1967USD} + trade_{baB\ t\ 1967\ USD}}$).

[34] The version of the IMF DOTS database that we are working with includes information for a total of 1’344’648 unidirectional trade flows. 808,301 of them have two mirror entries, but only 518,517 are recorded independently (with the remainder being imputed from partner records by the IMF). The latter figure corresponds to 38.6 percent of all observations. In terms of volume, they account for 78.0 percent of total trade.

Monadic ABBA term for country A's imports in year t:

$$\left|\sum_{i=1}^{n} trade_{baA\,t} - \sum_{i=1}^{n} trade_{baB\,t}\right| \; i.i.f.trade_{ba} \; recorded \; twice \; independently$$

Monadic ABBA term for country A's exports in year t:

$$\left|\sum_{i=1}^{n} trade_{abA\,t} - \sum_{i=1}^{n} trade_{abB\,t}\right| \; i.i.f.trade_{ab} \; recorded \; twice \; independently$$

These separate measures are important, because many analyses explicitly focus on either imports or exports. That said, they can be fused to craft a total-trade monadic ABBA term, which can again be normalized, for example by relating it to total trade or GDP.

Figure 5 illustrates the *monadic ABBA terms* for the USA and China. Dark grey lines track the value of total annual merchandise imports [exports] according to official American [Chinese] statistics, divided by the respective economy's GDP (retrieved from the World Development Indicators database). Dashed lines add up those bilateral imports [exports] that have also been recorded independently by the partner economy, again as a share of the "home" country's GDP. Dotted lines do the same, but using trade partners' data. The monadic ABBA term then refers to the difference between the dashed and dotted lines. In the examples below, it is relatively small for the USA, generally below one percentage point of GDP. It is larger for China, where it frequently exceeds five percentage points. Considering that part of the errors in the dyadic terms should cancel each other out in the aggregated, monadic figures, such gaps are remarkable.

Figure 5. Graphical illustration of monadic ABBA terms for USA and China

USA *China*

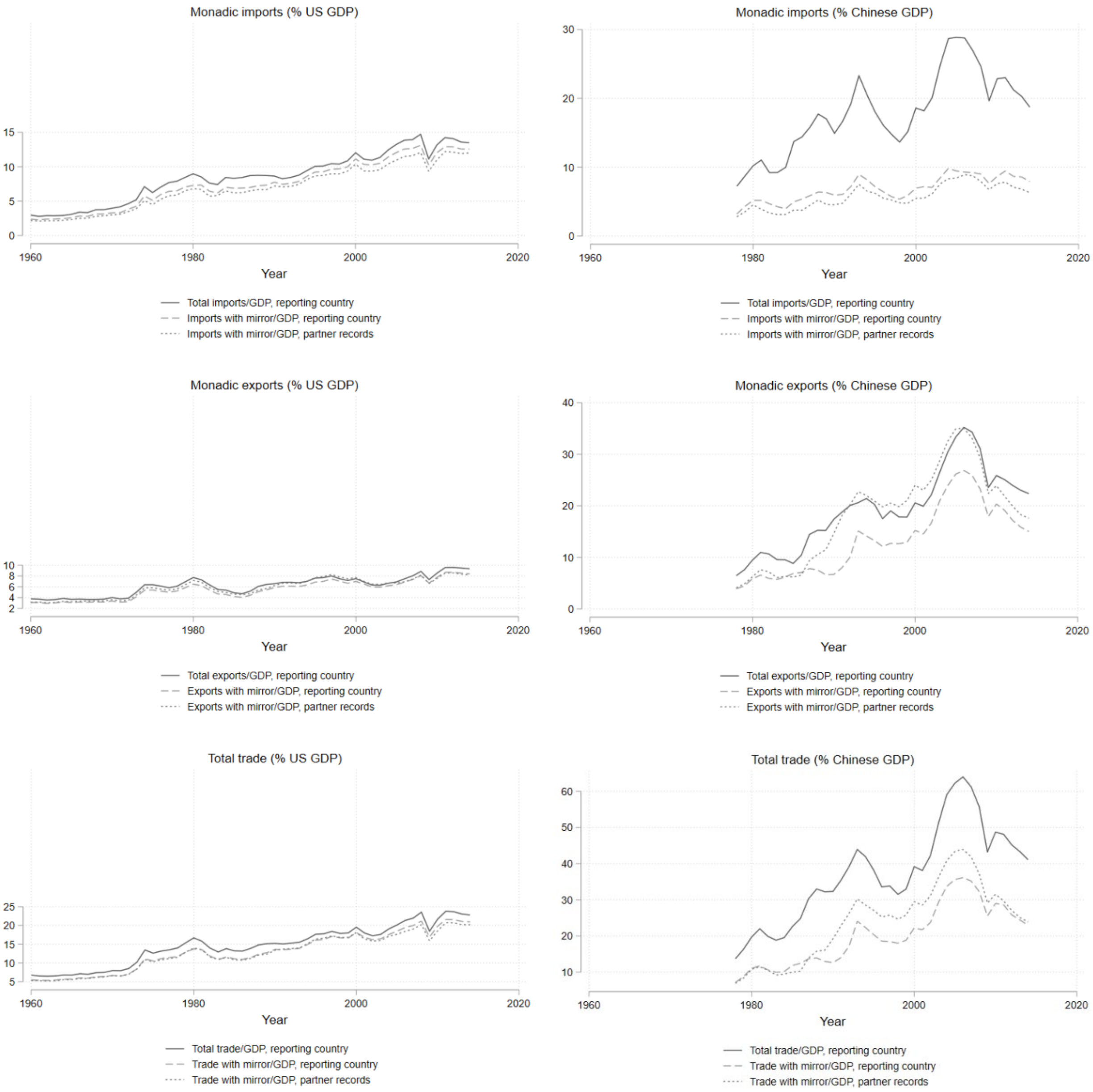
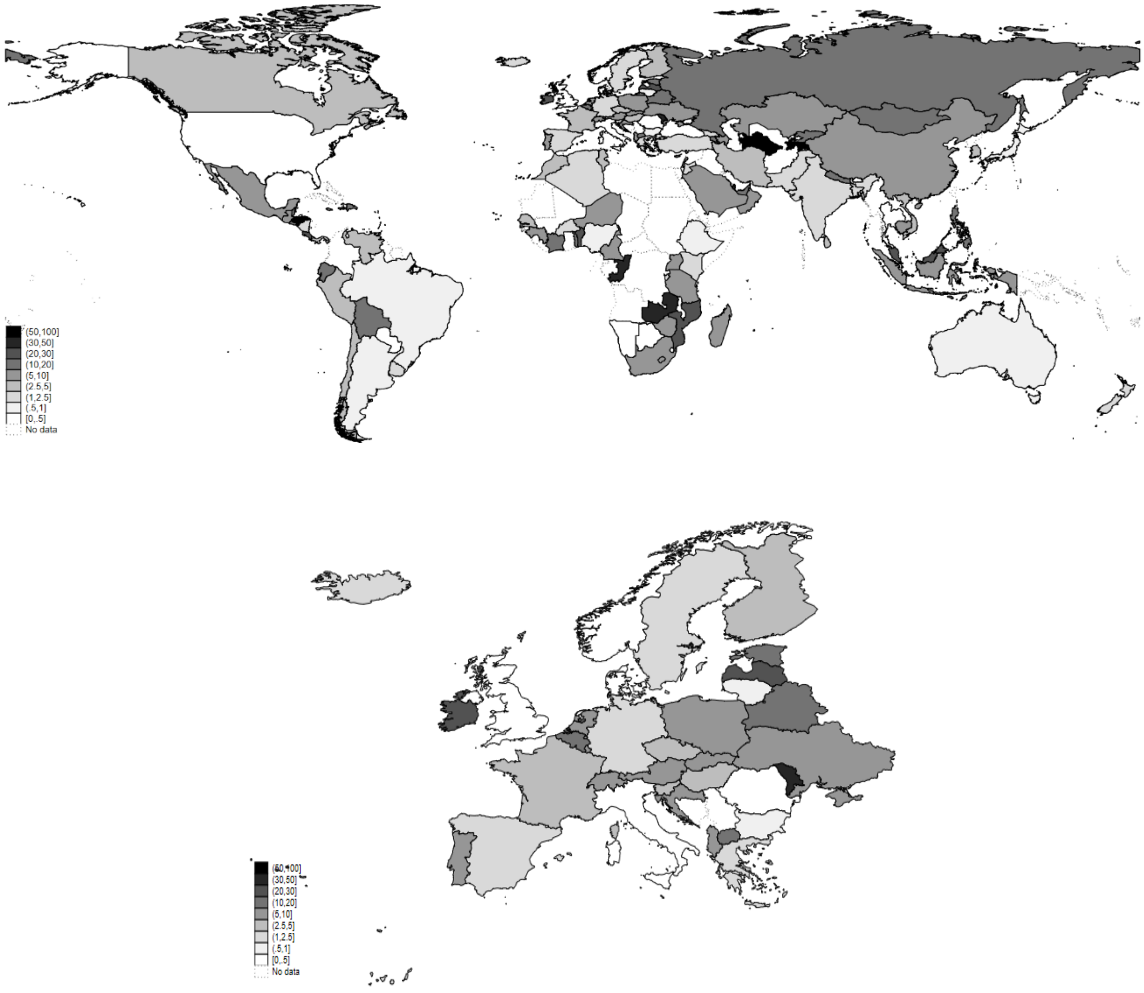


Figure 6 shows the global distribution of *monadic ABBA terms* for the year 2000. The darker the shading of the country, the larger is the potential measurement error. The upper part of the map shows the entire world; the lower part zooms in on Europe. Mirror discrepancies are not concentrated in any specific region of the world. Nor are there immediately obvious drivers of discrepancies. The monadic ABBA terms are very large in some Sub-Saharan countries, echoing the work of Jerven.[35] But they are also high in several G20 economies such as Russia, Canada, Mexico or Indonesia and advanced economies like Ireland, Belgium or Switzerland. Other years tell a similar story of widespread and seemingly random discrepancies (see appendix figure A2). In the global sample, the average monadic ABBA difference amounts to a sizable 7.6 percent of GDP (median 3.1 percent). For the decile with highest discrepancies it is above 18 percent. Discrepancies tend grow over time and are somewhat larger for more recent years.[36]

[35] Jerven 2013.

[36] In the 2000s-2010s subsample, the mean is 9.9 percent of GDP (median 5.1 percent).

Figure 6. Snapshots of monadic ABBA for total trade as share of GDP in the year 2000



SOURCE: own calculations based on IMF DOTS.

3.2. Uncertain origins of mirror discrepancies

How can we make sense of these discrepancies? And to what degree are ABBA terms, and hence data errors, distributed non-randomly? To the extent that the latter is the case, we are not only dealing with poor data, but with systematically skewed images of global trade.

There are plenty of potential explanations for the discrepancies. The first, commonly-cited one is “cost of insurance and freight” (c.i.f.). It is included in import prices, but not in the price of exports, which are loaded “free on board” (f.o.b.). Already in the 1950s, BOP statisticians debated how to tackle this issue, eventually settling for a flat-rate top-up.[37] According to a 1967 IMF report

> [balance] of payments folklore includes the notion that c.i.f. can be reduced to f.o.b. by deducting 9 or 10 per cent for freight and 1 per cent for insurance, and to act upon this convention may do no great violence to the balance of payments.[38]

That solution had its critics. John S. Smith, IMF assistant director of BOP statistics, lamented at the time that “the Fund staff has arbitrarily assumed in almost all cases that freight amounted to 9 per cent of the c.i.f. value. This percentage is believed to be somewhat on the high side.”[39]

Since then, falling trade costs have shrunk the c.i.f.-f.o.b. difference. Today, it is less than 2 percent of the value of US-EU and intra-European trade, and no more than 5-7 percent for trade between USA/EU and China/India.[40] Furthermore, the actual discrepancies often point in the opposite direction from what the c.i.f.-f.o.b. difference would suggest. In the US-Mexican example above, it is the net *exporter*, not the net *importer*, that consistently provides higher estimates of the trade surplus. In the USA and China monadic data in Figure 4, the direction of discrepancies is generally consistent with the c.i.f.-f.o.b. difference. But volatility on a year-on-year basis is not.

[37] Verbatim Report of the International Monetary Fund Meeting of Fund Statistical Correspondents on Balance of Payments Discussions held at the Burgundry Room - Sheraton-Park Hotel, Washington D.C. on Thursday, September 27, 1956 at 2:30 pm, retrieved from IMF Archives, Washington D.C.

[38] Alves 1967, 7, retrieved from IMF Archives, Washington D.C.

[39] Smith 1966, 14, retrieved from IMF Archives, Washington, D.C. Also Yeats 1978, 350.

[40] Miao and Fortanier 2017.

Limited statistical capacities are a second driver of mirror discrepancies.[41] Some countries have better resourced data collection systems than others, and economic crises or wars can undermine data collection.[42] That does not mean that rich country statistics are necessarily accurate, but simply that imperfect data collection is an inevitable source of disagreements, and that ceteris paribus it is particularly pronounced where statistical capacity is limited.

Accounting-technical glitches or cross-country differences in statistical practices are a third source of discrepancies.[43] Even when countries use similar concepts, classification systems can differ. Countries can attribute the same flow to different accounting periods, potentially affecting recorded values if exchange rates fluctuate.

Fourth, trading entities face incentives to misreport the value of shipped goods. High tariff rates encourage the under-invoicing of imports,[44] export subsidies the over-invoicing of exports.[45] Over-invoiced imports are used to move money abroad, circumventing capital controls.[46] EU common market rules encourage over-invoicing exports to evade VAT payments.[47] Smuggling and other illicit trade escapes data collection altogether. And sometimes governments deliberately occlude sensitive transactions, such as arms trade.

Globalizing production is a final driver of discrepancies.[48] Trade in intermediate goods and merchanting causes conflicting attributions of source and destination countries.[49] The USA may

[41] Jerven 2013.
[42] Schultz 2015.
[43] International Monetary Fund 1993.
[44] Bhagwati 1964.
[45] Bhagwati 1967.
[46] Yeats 1990.
[47] Braml and Felbermayr 2019.
[48] UNECE, Eurostat, and OECD 2011; Linsi and Mügge 2019.
[49] Interview with senior trade statistician at OECD Statistics Directorate, Paris, 6 June 2017. Note that the issue was already recognized as a problem back in the 1950s; see Verbatim Report of the International Monetary Fund Meeting of Fund Statistical Correspondents on Balance of Payments Discussions held at the Burgundry Room - Sheraton-Park Hotel, Washington D.C. on Thursday, September 27, 1956 at 2:30 pm.

record Chinese goods that arrived via Singapore as imports from China. But the Chinese might register an export to Singapore (not to the USA). And the Singaporeans note an export to the USA, while American figures record no import from there. Such mismatched reporting is still compliant with the guidelines in *International Merchandise Trade Statistics: Concepts and Definitions*,[50] given that ultimate destinations may be unknown to the exporter.[51] On top, slanted transfer pricing by multinationals,[52] unclear classification of export processing zones,[53] and low-tax or secrecy jurisdictions aggravate reporting problems.

It is a challenge to attribute observed discrepancies to these specific drivers. Some of them are difficult to observe or proxy. And plenty of proxies also pull in different directions. For instance, advanced economies can be expected to have high statistical capacity, but are also deeply integrated into complex global value chains that cloud our view on trade flows. That applies *a fortiori* to members of the EU single market—which not only features highly integrated and geographically distributed production, but is also not covered by regular customs controls.

The ambiguities in empirically exploring such drivers leave us with the worst of two worlds. On the one hand, plausible data distortions vary systematically: are trading nations rich or poor? Are large multinationals domiciled there? Do they offer tax advantages to attract corporate activity, at least on paper? Are they global trade hubs? On the other hand, because these factors blur our data *simultaneously*, disentangling them is nigh impossible—as noted by Yeats[54] and consistently raised in interviews with OECD, WTO and IMF statisticians.[55]

[50] United Nations Statistics Division 2011.
[51] Markhonko 2014.
[52] Ylönen and Teivanen 2018.
[53] Markhonko 2014.
[54] Yeats 1990, 136–137.
[55] Interview with senior trade statistician at OECD Statistics Directorate, Paris, 6 June 2017; interview with senior WTO statistician, Geneva, 22 August 2017; interview with IMF statisticians, Washington D.C., 19 September 2017.

This said, a number of variables *can* be well measured or proxied, allowing some exploration of the extent to which they drive mirror discrepancies. To provide such exploration, we have analyzed our most fine-grained discrepancy data: the dyadic ABBA terms for a substantial cross-section of dyads and more than half a century (1950-2004).[56] Our dependent variable builds on the elemental ABBA term introduced above. It takes the log value of the absolute ABBA discrepancy for a dyad-year divided by the sum of the two mirror flows, all in constant 1967 USD.[57]

We compare the model fit and the explained variance for a range of specifications. Model 1 controls for mirror-average trade volume, the average of dyad-specific c.i.f. conversion rates computed by the OECD,[58] and a dummy equal to 1 if both countries in a dyad are non-OECD economies to flexibly evaluate the role of economic development. Model 2 also controls for similarity or closeness of dyads in geographic, political and cultural terms, their level of economic development, EU membership and democracy, while avoiding multicollinearity issues (to check multicollinearity we rely on variance inflation factors). We also include a dummy for all trade flows involving at least one oil export-dependent economy,[59] as well as those involving five well-known entrepot trade jurisdictions,[60] and a dummy for China, whose data is frequently brandished

[56] We rely for this analysis primarily on the Tomz dataset on trade flows, a dataset that is itself based on IMF DOTS, covering thousands of dyads over a 50-plus year period, and including a broad range of explanatory and control variables that are relevant for our investigation. We add information on mirror trade flows, which we derive from IMF DOTS, as well as some additional explanatory variables to that dataset.

[57] Formally, $\log\left(\frac{|trade_{baA\,t\,1967USD} - trade_{baB\,t\,1967USD})|}{trade_{baA\,t\,1967USD} + trade_{baB\,t\,1967\,USD}}\right)$. In robustness tests we use a variety of other operationalizations of the DV. These yield very similar results (see appendix tables A2 to A7).

[58] The conversion rates are based on the work described in Miao and Fortanier (2017). Combining explicit c.i.f.-f.o.b.. rates and gravity model estimates, they estimate product-level transport and insurance costs for each dyad-year for the period from 1995-2014. We use a dataset (provided by the authors) with product-weighted dyad-level annual c.i.f. rates. We calculate the 1995-2014 period averages for each dyad, which we treat as the (constant) dyad-level "best guess" for c.i.f. rates throughout our longer time period.

[59] Iraq, Libya, Venezuela, Algeria, Kuwait, Azerbaijan, Sudan, Nigeria, Saudi Arabia, Oman, Kazakhstan, Russia, and Iran.

[60] Singapore, Panama, United Arab Emirates, Netherlands, and Belgium.

as unreliable. Model 3 includes year-fixed effects; Model 4 adds dyad-fixed effects to those. In separate analyses (appendix table A1) we also examine the role of applied tariff rates and capital account openness, for which measures are available only for a limited subset of our sample.

Table 1 below summarizes the main results. Unsurprisingly, a trade flow's size is a powerful predictor of the absolute size of a discrepancy. Dyads of less developed states tend to have larger discrepancies than developed ones, and, as already indicated in Figure 5, relative discrepancies are smaller for dyads that trade more with each other. Notably, c.i.f. conversion rates *per se* do not appear to drive asymmetries significantly. Model 3 shows that countries that are further removed from one another geographically, culturally and politically tend to report higher discrepancies relative to trade volume. The same is true for dyads involving island-states, landlocked states and countries with large territories (and many border checkpoints). In line with previous studies, we find at least weak evidence that more democratic countries and dyads yield slightly smaller mirror discrepancies.[61] GATT/WTO membership and preferential trade agreements correspond to smaller discrepancies in between-effect models. Entrepot and oil trade is associated with higher discrepancies. The China dummy is not significant.[62]

Counterintuitively, controlling for all other factors, higher estimated c.i.f. rates are associated with smaller discrepancies, and EU membership is consistently related to higher discrepancies—a point we take up below. A number of these substantive results disappear once full dyad and year fixed effects are included. And not surprisingly, measures of model performance, such as Akaike's and Schwarz's Bayesian information criteria (AIC and BIC), suggest that successive inclusion of control parameters improve model performance, with the full

[61] Hollyer, Rosendorff, and Vreeland 2011.

[62] Trade through special administrative regions is not recorded separately in IMF data, which can alleviate distortions due to re-exports through Hong Kong described in other studies.

fixed effects model 4 performing best. The complementary analysis in appendix table A1 suggests that in the subsample of observations for which the information is available (restricted to 1988-2004), greater capital openness is associated with somewhat smaller discrepancies, and higher tariff rates with larger discrepancies, in line with expectations of deliberate over- and under-invoicing. But these relationships are statistically insignificant when including other controls.

The most striking result is, in fact, how little variation the various explanatory variables explain-away—even in the full fixed effects model (model 4). The size of trade flows does most of the explanatory work—which is neither surprising nor particularly elucidating. Taken together, absolute trade volumes and dummies for non-OECD economies together account for (only) 22 percent of variation. And adding all other variables, or year-fixed effects, barely improves model fit (see 0.24 R-squares in Models 2 and 3). Including full dyad-fixed effects and all relevant substantive parameters still explains less than half of variation (resulting R-square of 0.47).[63]

[63] The conclusion is similar if one considers measures of model fit, such as AIC and BIC.

Table 1. Sources of ABBA-measured mirror discrepancies

DV: Mirror discrepancy relative to sum of mirror flows (log, constant USD)	(1)	(2)	(3)	(4)
Trade volume	-0.24 (-64.08)	-0.22 (-41.14)	-0.21 (-40.18)	-0.20 (-34.66)
C.I.F. rate (dyad mean)	0.03 (0.06)	-1.62 (-2.80)	-1.95 (-3.32)	
Distance		0.10 (7.05)	0.11 (7.24)	
Shared border		0.12 (2.02)	0.11 (1.91)	
Number of landlocked in dyad		0.10 (5.56)	0.08 (4.64)	
Number of island-states in dyad		0.15 (5.91)	0.08 (4.64)	
Land area (product)		0.03 (6.45)	0.04 (7.22)	
GDP (product)		-0.01 (-2.04)	-0.03 (-3.53)	-0.03 (-2.64)
Both industrial states		-0.20 (-4.87)	-0.17 (-3.91)	
Both non-industrial states	0.26 (12.93)	0.27 (12.81)	0.25 (11.41)	
Polity IV score (product)		-0.02 (-2.30)	-0.19 (-2.28)	0.001 (0.22)
Both formal GATT/WTO members		-0.06 (-3.65)	-0.07 (-4.21)	0.02 (1.38)
Reciprocal PTA in force		-0.20 (-8.60)	-0.21 (-8.94)	-0.13 (-5.78)
Common currency		0.19 (2.71)	0.16 (2.63)	-0.11 (-1.25)
Both EU members		0.50 (6.27)	0.49 (6.03)	0.28 (3.83)
Common colonial orbit		-0.43 (-3.56)	-0.41 (-3.43)	
Common language		-0.10 (-3.41)	-0.10 (-3.51)	
Oil exporter		0.10 (3.87)	0.09 (3.52)	
Entrepot trade hub		0.15 (4.12)	0.15 (4.18)	
China dummy		0.02 (0.39)	0.003 (0.07)	
Year-fixed effects?	No	No	Yes	Yes
Dyad-fixed effects?	No	No	No	Yes

(continues on next page)

Number dyads	9,689	9,689	9,689	9,689
N	184,426	184,426	184,426	184,426
R^2	0.22	0.24	0.24	0.47
AIC	587151.1	582370.7	582003.5	514120.4
BIC	587191.6	582583.4	582216.2	514738.0

NOTE: T-statistic in parentheses. Dyad-clustered robust standard errors. Constant omitted from output.

Additional analyses yield comparable results. We have estimated similar models using the dyadic ABBA discrepancy relative to the value of the import-based record (appendix table A2), the mirror average (appendix table A3), a non-logged version of the DV (appendix table A4), as well as the ABBA value in absolute terms (rather than relative to trade volume), both in 1967 USD (appendix table A5) and in current USD (appendix table A6). In a further check, we have dropped dyads in which at least one trade flow has a value of zero (appendix table A7). Doing so actually decreased R-square to 0.12 in the model corresponding to Model 1.

These analyses buttress the qualitative finding that the sources of discrepancies are highly idiosyncratic, hard to determine in any particular instance and, as a result, difficult to control for in empirical analyses. We do not know to what degree the ABBA terms reflect multiple, layered biases versus unsystematic error. In any case, we cannot assume that errors are randomly distributed and therefore cancel each other out at the aggregate level. As our replication analyses below show, export records-based data at times generates *stronger* statistical results than import records-based one (even if at times in an opposite direction). Export-based trade values are not just a noisy version of import data, which would simply introduce measurement error and weaken import-based results due to attenuation bias. Instead, there are systematic differences between the two. That makes it imperative to heed these differences in empirical analyses involving trade.

3.4. The handling of the mirror problem in existing IR and IPE scholarship

Several datasets have been developed to address the mirror problem, including the GTAP[64], BACI[65] and OECD BIMTS[66] databases. Notwithstanding minor differences in methodology,[67] they all try to "balance" mirror flows through weighting by reporter reliability, inferred from the size of a reporting economy's discrepancies with the data from all other countries. However, none of these databases have been designed with an IR/IPE user base in mind: they cover only subsamples of countries and short time periods.[68] And they are designed primarily for dyadic-product-level analyses rather than country-dyads. These data bases, including BACI, therefore cannot readily be used to answer typical IR/IPE research questions. Only one of the 108 papers that we reviewed above can and does use one of those datasets (cf. Table 1).[69] The methods developed in these databases, furthermore, cannot be fully extended to other country-dyads, years or products, since the computer codes used to generate inferred reporter reliability are not publicly available. Below, we therefore develop our own approach to such balancing to the extent that current data allows.

Even though IMF DOTS provides both sides of the mirror data, almost all studies we have reviewed use the import values, either consciously or by using the major off-the-shelf datasets in IR and IPE (e.g. COW or Gleditsch). Values from partner countries' export statistics are disregarded. Researchers frequently justify this practice arguing that authorities have greater

[64] Gehlhar 1996.
[65] Gaulier and Zignago 2010.
[66] Fortanier and Sarrazin 2016.
[67] An useful overview is provided in Ibid.
[68] GTAP's most recent release (GTAP 10) includes data for 121 countries for four reference years (2004, 2007, 2011, 2014); BACI's 2020 update covers 200 countries for the period 1994-2018; OECD BIMTS is work in progress that feeds into the TiVA initiative, which currently encompasses data on 120 countries between 2007-2016.
[69] The study by Osgood (2017), which uses BACI for part of the analyses.

incentives to monitor imports than exports for the collection of customs duties. Ceteris paribus import data should be better.

This justification for ignoring export values is not convincing. A number of factors may, at least at times, argue in favor of export statistics. First, importers have greater incentives than exporters to distort the declared value and ultimate origin or destination of merchandise, because they, not the exporters, normally pay custom duties. Second, when exporting countries have higher statistical capacity than importers, their records are likely to be more accurate, as well. Third, inside custom unions—such as the European Union, which accounts for roughly 15 percent of global trade and is free of internal custom inspections—governments rely primarily on value-added tax (VAT) data to estimate trade flows.[70] Since sellers (ie, exporters) pay VAT, exporting-country records will be more reliable in such cases.[71] Fourth, growing e-commerce and disintermediation mean that such dynamics affect global trade as well. As private consumers increasingly buy products online from providers abroad, import statistics will miss growing shares of global trade, while exporters have to meet more stringent declaration obligations.[72]

A priori, then, we have no reason to assume that one set of figures is invariably superior to the other. A senior OECD statistician highlighted this point in an interview:

> When [academic researchers] have … tried to resolve asymmetries … they said "let's just look at imports and forget about exports and then you define asymmetries away"… that's nice if they're small, but it doesn't really

[70] Eurostat 2016.

[71] As our analysis of trade statistics discrepancies in the EU-27 sample in appendix table A8 shows, within-dyad discrepancies overall tended to increase as European countries became member of the Common Market – an effect driven by deterioration in import records (Model 4).

[72] Braml and Felbermayr 2019.

> work well in total. […] Discrepancies are large. You can't say it's a rounding error.[73]

3.5. Suggested approaches to better account for the mirror problem in IR and IPE research

The mirror problem operates at various levels, and there is no one way to solve it. But there are ways to better account for it in our analyses, which we demonstrate in the replication exercises below. For in-depth analyses of specific datapoints (e.g. say, the US-Mexico trade imbalance in 2017) it may be possible to use priors and triangulation to explain the asymmetries and determine a plausible range of values. For large-n analyses that is not an option. But it is possible—and in our view necessary—to check the *robustness* of trade-related findings to measurement problems. That includes sensitivity of results to missing values, as discussed above.[74] The other dimension, and our focus here, concerns the mirror problem.

In dyadic setups, mirror records can be leveraged in several, complementary ways. If we want to avoid strong assumptions about the sources of measurement errors, one option is to run, and to consider as equally (in)valid, analyses using either side of the mirror (the "mirror substitution check"). If a finding holds after substitution, it suggests that the results are not driven by measurement artefacts. If it does not, we need to investigate further. Alternatively, we can use weighted averages of mirror records, as BACI for example does, a strategy we prefer over exclusive reliance on one side of the mirror.

The next section outlines two ways to implement a mirror-weighting (a simple weighting, and a residuals-based one) for the datasets most commonly used in IR and IPE. Such approaches

[73] Interview with senior trade statistician at OECD Statistics Directorate, Paris, 6 June 2017.
[74] Boehmer, Jungblut, and Stoll 2011; Barbieri and Keshk 2011.

have their own shortcomings. Given the uncertain origins of mirror discrepancies, no weighting can reveal true trade values. But in combination, weighted averages and mirror substitution checks help to indicate the plausible size of trade-related coefficients and their robustness. Furthermore, while the mirror problem is a dyadic phenomenon, the information captured in the ABBA terms can support monadic analyses as well. One approach, further explored below, uses discrepancies between the sums of available mirror records to gauge how reliable monadic trade data of a given country-year is. Of course, the ABBA information largely ignored in IR and IPE research could also be leveraged in other ways. It could, for instance, help select appropriate measurement error models,[75] serve as an input for measurement error-injection tests,[76] or as a basis for identifying major measurement outliers to be considered and contextualized in more qualitative exploration. Below, however, we concentrate on the versatile and easily implementable approaches to directly assess the sensitivity of key results to explicit gauging of the mirror measurement problem.

4. The mirror problem in IR and IPE studies of trade: five replication analyses

This section re-evaluates prominent studies about economic interdependence. We probe how sensitive the findings are to data uncertainty, and we explore avenues to heed it better. Full replications of a large sample of IR and IPE studies exceeds the scope of this article. The five we have chosen have clear policy-relevance and stand prominently in the literature. They cover diverse uses of trade data and research designs, and cover both IR and IPE, broadly construed, and dyadic as well as monadic set-ups (Table 2).

[75] Carroll et al. 2006.

[76] Neumayer and Plümper 2017.

Table 2. Selection of studies for replication

	International Security/Politics	**Political Economy**
Dyadic	Kastner (2014) (effect of trade on security diplomacy)	Rose (2004)/Goldstein, Rivers & Tomz (2007) (GATT/WTO membership affecting trade)
Monadic	Barbieri & Reuveny (2005) (trade affecting violent conflict)	Garrett & Mitchell (2001) (trade affecting welfare states)

We first replicate the original findings, and then compare these to estimation approaches that explicitly consider the mirror problem. How we do so differs per study – particularly between the dyadic versus the monadic set-ups. In dyadic studies we can conduct a "mirror substitution check" head-on, comparing the results based on import and export data respectively. In a second step, we use weighted averages to estimate dyadic trade flows, explicitly quantifying the inferred reliability of import-based and export-based trade values. Weighted averages take the following basic form:

$$trade_{ba\,t\,wgt} = w_{a\,t} * trade_{baA\,t\,f.o.b.} + (1 - w_{a\,t}) * trade_{baB\,t\,f.o.b.}$$

We first convert imports into *approximate* f.o.b. values by deducting the mean dyad-specific c.i.f. rates that we estimate based on data provided by the OECD (these are mostly generated through a gravity model rather than observed and, for our purposes, treated as constant over time, cf. footnote 59). $w_{a\,t}$ is determined by the median of country A's ABBA discrepancies relative to the combined sum of mirror flows, $median\left(\frac{|trade_{baA\,t\,f.o.b.} - trade_{baB\,t\,f.o.b.}|}{trade_{baA\,t\,f.o.b.} + trade_{baB\,t\,f.o.b.}}\right)$, in its trade flows with all other countries in a specific year relative to that of partner country B (a value naturally bounded between 0 and 1). The smaller [larger] country A's median ABBA relative to that of country B, the higher [lower] the weight assigned to its reported intra-dyadic trade volume. Specifically,

$$w_{a\,t} = 0.5 + \frac{|ABBAmedian_{A\,t} - ABBAmedian_{B\,t}|}{2} \text{ if } ABBAmedian_{A\,t} \leq \; ABBAmedian_{B\,t};$$

$$w_{a\,t} = 0.5 - \frac{|ABBAmedian_{A\,t} - ABBAmedian_{B\,t}|}{2} \text{ if } ABBAmedian_{A\,t} > ABBAmedian_{B\,t}.$$

An alternative approach weights the reporting country's estimates not by its *full* average discrepancies but by the *unexplained* discrepancies, gleaned from the residuals of models estimating such discrepancies.[77]

In monadic settings, the mirror problem is harder to track because the source data does not directly offer mirror values. Still, since monadic data is central to many IR and IPE analyses, we propose easily implementable ways at least to gauge how mirror discrepancies may affect estimates. One approach is to collapse the weighted dyadic averages into monadic measures; however, the limited share of dyadic trade flows with two independent mirror entries (cf. footnote 35) constrains the usefulness of this approach to evaluate the effects of total monadic levels of trade. Therefore we outline two complementary approaches that can better accommodate changes in total monadic trade levels: including the monadic ABBA terms defined above as control variables, and visualizing the interaction between key explanatory variables and ABBA-proxied measurement uncertainty.

As supplements to this paper, we make available two public datasets of dyadic and monadic ABBA terms for a large panel of countries between 1950 and 2014, together with the code used to generate them as well as the weighted averages. Applied to our new datasets or any other dyadic ones it can enable researchers to adapt trade data to whatever context suits their research aims.

[77] We follow a similar procedure as for the simple weighted averages, but use different metrics to determine the weights: we generate the predicted values of a dyad-year's ABBA value in a model with the full set of explanatory variables and no fixed effects (specifically model 2 in appendix table A5). For each country-year we then calculate the share of observed trade values that have a smaller than predicted discrepancy. The higher that share, the higher the weight accorded to the reporter in question.

4.1. Replication of dyadic studies

Our first replication concerns a research design investigating what trade with China implies for geopolitical alliances. Our second re-examines analyses that link GATT/WTO membership to trade flows. In both instances, we follow the same steps: we replicate the original results (Models 1); re-run the baseline for the subset of the sample for which two independent mirror records are available (Models 2); replace the import-based records with the corresponding entries in export-based records (the "mirror substitution check"; Models 3); and replace trade values with the simple weighted average of mirror records (Models 4).

4.1.1 Kastner (2016)

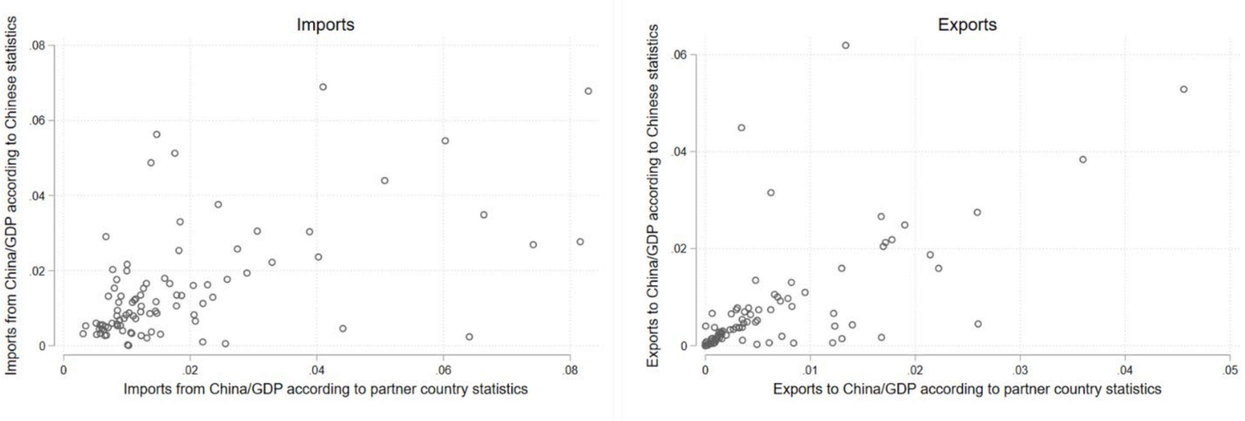
Kastner's *Journal of Conflict Resolution* study evaluates how countries' bilateral trade with China influences geopolitical alignments. Kastner tracked foreign government's support for three controversial moves by the Chinese government: the 2005 Anti-Secession Law opposing Taiwanese independence, the 2008 crack-down in Tibet, and seeking other WTO members' recognition as a market economy from 2004 onwards. He then analyzes bivariate correlations between the level of support and various measures of bilateral trade.

We first illustrate the mirror problem in the independent variable used in this setup. The left-hand scatterplot in Figure 7 compares mirror values for China's trading partner imports as a share of their GDP (Chinese import figures on the y-axis; partner country import figures on the x-axis). The right-hand plot does the same for exports. Both values are from 2004, the year before the declaration of the Anti-Secession Law—the specific case that we re-analyze.[78] The graphs

[78] Replication results are similar for the other two issue areas (Tibet and WTO market economy status). For reasons of space, we present only one of these three complementary analyses.

show that mirror discrepancies can be large, even when normalized by GDP. The correlations for import mirror records are 0.71 on a linear scale and 0.52 for its log transformation (relevant for the following analysis); they stand at 0.96 (linear) and 0.82 (logged) for exports.

Figure 7. Value of bilateral trade flows with China as a share of GDP in mirror statistics, 2004



NOTE: Observations for which the IMF indicates partner records imputations are excluded. For better readability both axes in both graphs are truncated at 0.1.

Kastner's original model is a cross-sectional multinomial logit.[79] The dependent variable is foreign governments' support for the Anti-Secession Law, coded into three categories: no, moderate, or strong support. The quantity of interest is the strength of the correlation with various measures of trade dependence, controlling for geographic distance, measures of authoritarianism, security relations with the USA, and national power. All data is from 2004. Trade dependence is operationalized as the value of foreign governments' bilateral imports from [exports to] China as a share of GDP, as well as their value relative to total imports [exports].[80] For both import and export values, Kastner relies on Chinese data: other countries' exports to China correspond to

[79] Kastner 2016, 992–994.

[80] We here only show results for the trade/GDP ratios. Results are very similar for measures of trade dependence relative to total trade.

imports from that country in Chinese records; the value of other countries' imports from China are derived from what China reports exporting to them.

Table 3 summarizes the imports-based analyses (full results in appendix table A9). Model 1 re-establishes the original results. In Model 2 we restrict the sample to those observations for which two independent mirror records are reported in IMF DOTS. Models 3 and 4 perform the ABBA robustness checks described above.

Table 3. Replication of Kastner (2016)

DV: Support Anti-Secession Law	(1)	(2)	(3)	(4)
Model	*Original Baseline*	*Baseline in sample with two independent mirror records*	*Mirror substitution check*	*Weighted mirror average*
Side of mirror	Chinese	Chinese	Partner countries	Average
Moderate support				
Imports from China/GDP (ln)	1.20 (3.56)	0.78 (2.61)	1.96 (3.42)	1.46 (3.17)
Strong support				
Imports from China/GDP (ln)	0.82 (3.41)	0.15 (0.58)	1.28 (2.42)	0.55 (1.16)
Control variables as in original?	Yes	Yes	Yes	Yes
N	146	96	96	80
Log-PLH	-105.3	-65.5	-61.7	-52.2

NOTES: No support is the base outcome; robust standard errors; z-statistic in parentheses.

The robustness tests strengthen but also nuance the original findings. Using the same Chinese data as the original study but in the reduced sample for which we have independent mirror records, the coefficients are notably (and understandably) smaller and lose significance for strong support. Using partner-country records for the same sample, the correlation coefficients jump from

0.78 to 1.96 for moderate and from 0.15 to 1.28 for strong support of the Anti-Secession Law (cf. model 3 vs. 2).

These differences are substantively meaningful: model 2 predicts the probability of a country expressing no support for the Anti-Secession Law to decrease from 51 percent (at the 25th percentile of trade dependence) to 40 percent (at the 75th percentile). Model 3, in contrast, predicts a much bigger decrease: from 63 to 30 percent.[81] Using weighted averages of import measures also strengthens the relationships compared to the Chinese data.[82] Compared to the original results, these findings underline how import dependence amplifies moderate support for China's Anti-Secession Law, while the increase in strong support is less conclusive.

In this instance, the original findings "pass" the ABBA robustness checks, strengthening our confidence in the positive correlation found in the original study. The replication is also informative in light of possible publication bias. The magnitude and strength of the correlations have shifted substantially. Many past analyses are likely to have produced statistically insignificant results using one side of the mirror, while the other side or a weighted average might well have generated statistically significant (and thus more obviously publishable) findings. The mirror problem thus affects not only what we think to know about the origins and effects of trade, but also what we think to know *not* to be the case (type II error).

[81] The probability of moderate support increases from 9 to 20 percent in model 2 vs. 7 to 22 percent in model 3. The probability of strong support is virtually unaffected according to the data in model 2, but jumps from 30 to 48 in model 3. All marginal effects are calculated holding all other variables at their median values.

[82] The results are similar for the residuals-based weighted averages, see appendix table A9. The reduction in sample size from Model 3 to 4 is due to missing data in OECD's c.i.f.-f.o.b. conversion rates, which we use to calculate weighted averages.

4.1.2. Rose (2004)/Goldstein, Rivers and Tomz (2007)

Our second set of replications assesses an IPE setup with a large time-series dataset with global coverage and trade as the dependent variable. We examine two prominent studies with contrasting conclusions about the trade-facilitating effects of the GATT/WTO. A much-cited article by Rose found no positive—and in some models actually a negative—effect of GATT/WTO membership on bilateral trade volumes.[83] Goldstein, Rivers and Tomz (henceforth GRT) later challenged this result.[84] The disagreement centered on two issues: Rose conducted a cross-sectional analyses, focusing on "between" unit variation; GRT analyzed within-effects, so variation over time within a given unit. In addition, Rose classified country membership by formal participation in GATT/WTO; GRT considered a more fine-grained categorization, accounting for the de facto but not de jure ("informal") participation in the regime by some countries, especially (former) colonies.

We use the dataset provided by Tomz and add the mirror information from the IMF DOTS database.[85] We drop the 83,346 observations which are either missing in our dataset or outliers for which the log-difference in import-based records is greater than 1, generating 298,310 dyad-year observations. For 77,354 of these, IMF DOTS gives no mirror record, and for 37,309 the IMF has used partner records to impute missing values. This leaves us with 183,647 dyad-years with two independently recorded values. The dependent variable used in the analyses is the log of the value of bilateral trade flows in 1967 US dollars.

[83] Rose 2004.

[84] Goldstein, Rivers, and Tomz 2007.

[85] Although GRT also rely on IMF DOTS as their main data source, they complement it with some hand-collected data points and the imputation of missing values from partner records.

We first replicate Rose's between-effects model (Table 4; full results in appendix table A11), then GRT's within-analysis (Table 5; full results in appendix table A12). Models 1 and 2 again re-establish the original results and repeat the analysis with the restricted sample of dyads with independent mirror records. Models 3 and 4 perform the ABBA checks.

In this setup, as alluded to in Linsi and Mügge,[86] the implications of the mirror problem are stark. The import figure sub-sample with two independent mirror records corroborates Rose's negative relationship between formal GATT/WTO membership and bilateral trade (Model 2 in Table 4). But the coefficients become strongly *positive* and statistically significant once we use the corresponding export figures (Model 3 in Table 4). Formal GATT/WTO members appear to trade *less* than non-members if we use import-records, but they trade *more* if we use export figures. If we plug in the weighted averages data, the effect becomes negative for dyads in which both countries are formal GATT/WTO members, while the coefficient for one formal member is smaller and insignificant at the five-percent level.

[86] Linsi and Mügge 2019, 370.

Table 4. Replication of Rose (2004)

DV: bilateral trade	(1)	(2)	(3)	(4)
Model	*Original Baseline* (excl. large differences DOTS vs. GRT)	*Baseline in sample with two independent mirror records*	*Mirror substitution check*	*Simple weighted mirror average*
Side of mirror	Import-records	Import-records	Export-records	Average
Both formal members	-0.10 (-3.16)	-0.15 (-3.88)	0.56 (5.53)	-0.13 (-3.22)
One formal member	-0.20 (-6.58)	-0.16 (-4.31)	0.48 (4.61)	-0.07 (-1.71)
Control variables as in original?	Yes	Yes	Yes	Yes
Year-fixed effects?	Yes	Yes	Yes	Yes
Dyad-fixed effects?	No	No	No	No
Years	1950-2004	1950-2004	1950-2004	1950 -2004
N	298,310	183,647	183,647	177,473
Dyads	15,120	9,842	9,842	9,299
R^2	0.62	0.67	0.39	0.69

NOTE: Robust standard errors clustered by dyad; t-statistic in parentheses.

The results are equally remarkable for the GRT replications. They corroborate the GRT claim of a positive GATT/WTO membership effect. The mirror substitution check (Model 3 vs. 2 in Table 5) shows this effect to be *several times larger* once we use export-based data. Import data suggests a 30-35 percent boost to bilateral trade from GATT/WTO membership; export data puts the figure at 150-250 percent.

Table 5. Replication of Goldstein, Rivers and Tomz (2007)

DV: bilateral trade	(1)	(2)	(3)	(4)
Model	*Original Baseline* (excl. large differences DOTS vs. GRT)	*Baseline in sample with two independent mirror records*	*Mirror substitution check*	*Simple weighted mirror average*
Side of mirror	Import-records	Import-records	Export-records	Average
Both formal members	0.35 (8.22)	0.26 (4.99)	1.31 (7.82)	0.34 (6.58)
One formal member	0.18 (4.73)	0.12 (2.47)	1.10 (7.01)	0.20 (4.25)
Formal and nonmember participant	0.36 (7.74)	0.28 (4.94)	0.96 (5.19)	0.28 (4.87)
Both nonmember participants	0.45 (4.48)	0.30 (2.24)	-0.10 (-0.18)	0.16 (1.11)
One nonmember participant	0.08 (1.54)	0.10 (1.47)	0.49 (2.25)	0.23 (1.77)
Control variables as in original?	Yes	Yes	Yes	Yes
Year-fixed effects?	Yes	Yes	Yes	Yes
Dyad-fixed effects?	Yes	Yes	Yes	Yes
Years	1950-2004	1950-2004	1950-2004	1950-2004
N	298,310	183,647	183,647	177,473
Dyads	15,120	9,842	9,842	9,299
R^2	0.85	0.88	0.69	0.89

NOTE: Robust standard errors clustered by dyad; t-statistic in parentheses

Detailed data analysis shows these large differences to be driven by large discrepancies in a small number of dyads in which one country reports zero trade while the other does not.[87] Estimating the model with the two versions of weighted mirror averages, the results are substantively and statistically stronger than in the import-based baseline for formal GATT/WTO membership, but weaker for countries' accession to "informal" membership— the theoretical core of GRT's article.

In short, the mirror problem has important implications for this debate. Overall, attention to mirror discrepancies strengthens the trade-enhancing effect of formal GATT/WTO membership in statistical and substantive terms. This is good news from the GRT perspective and bad news for Rose's—irrespective of the other, originally-reported substantive and statistical disagreements—in terms of the overall effect. Also important, however, attention to the mirror discrepancies in trade data reveals that "informal" non-member participation appears to play a smaller role for the discrepant findings than previously estimated.

4.2. Extensions to Monadic Studies

The mirror problem is essentially a dyadic phenomenon. But mirror statistics can also be leveraged, even if imperfectly, to study data problems at the monadic level. The two replications we present illustrate easily implementable approaches to do so. They again cover different IR and IPE topics and research designs: the first study assesses how trade openness affects the risk of civil wars in developing countries; the second analyses the link between trade and government spending in advanced industrial economies.

[87] Note that dyads with missing values or mirror-imputed flows are excluded from the sample, so that these values refer to actually reported 0s.

Because coverage of bilateral trade flows is imperfect, we cannot perform full mirror substitution and average checks at the monadic level. We can aggregate weighted averages to the monadic level, but they capture only variations in subsets of a country's volumes of trade (with coverage fluctuating within and across countries). Therefore, our main replications privilege an alternative procedure that is better suited to assess measurement problems in reported total levels of monadic trade: we re-establish the original results (Models 1); re-run the baseline for the sample for which monadic ABBA terms are available (Models 2); include the monadic ABBA term as a "control" (Models 3); and re-run the baseline in a restricted sample that excludes the decile of observations with the largest mirror discrepancies (Models 4). Finally, we interact the monadic ABBA term with the explanatory trade-variable to visualize how measurement errors affect statistical findings.

Such sensitivity analyses alone cannot solve measurement problems. The ABBA term can "control" for measurement uncertainty, but it may also capture institutional dimensions of theoretical interest (e.g. economic development, economic structure or state capacity, to the degree that it correlates with these). Furthermore, correlation between the standard errors of the trade variable and the ABBA term biases the estimated coefficients. Dropping country-year observations with high ABBA terms can indicate the direction of bias, but it may also introduce selection problems. Thus, these checks do not aim to correct measurement errors *per se*, but to gauge how the latter may influence the statistical relationships of interest.

4.2.1 Barbieri and Reuveny (2005)

Monadic trade data is central to studies that link economic openness and the risk of civil wars. Barbieri and Reuveny have offered a systematic investigation.[88] Building upon Fearon and Laitin,[89] they assess various globalization measures (trade openness, FDI inflows, PFI inflows and internet usage) as predictors of civil war *onset* and *presence* (duration). They find that greater trade openness does not prevent civil war *onset*, but significantly reduces the duration of internal conflicts.

Thanks to data provided by the authors, we could reproduce the original results exactly.[90] Our replications estimate the effects of the trade openness variable once we take data problems into account. Table 6 summarizes the behavior of the trade variable; full results, which separately also show the results for weighted monadic terms, are in appendix table A13.

Re-establishing the authors' main result, Model 1 confirms the negative and near-significant (90-percent threshold) relationship between civil-war presence and total trade in goods and services as a share of GDP. Model 2 is similar but uses the trade openness measure we calculate from DOTS, excluding services trade. The negative relationship strengthens somewhat, as does the statistical significance. For our purposes, Model 2 is the baseline replication of Barbieri and Reuveny.

[88] Barbieri and Reuveny 2005.

[89] Fearon and Laitin 2003.

[90] Following the original research protocol, we run a logit model with a dummy variable identifying the presence or absence of civil war in any country-year. We include the authors' four measures of economic globalization and control for GDP per capita, population size, geographic variables (mountainous territory, noncontiguous states and oil reserves), measures of political instability, democratization, ethnic and religious fractionalization, as well as a variable counting the years of peace having elapsed since the last internal conflict, and cubic splines to address temporal dependence. Standard errors are robust and clustered by country.

Table 6. Replication of Barbieri and Reuveny (2005)

DV: Civil War presence	(1)	(2)	(3)	(4)
Model	*Original baseline*	*Baseline merchandise trade*	*Monadic ABBA as control*	*Censoring ABBA top decile*
Total trade/GDP (t-1)	-0.013 (-1.64)			
Merchandise trade/GDP (t-1)		-0.015 (-2.33)	-0.009 (-0.90)	-0.010 (-0.81)
Monadic ABBA term (t-1)			-0.04 (0.03)	
All other variables of original model included?	Yes	Yes	Yes	Yes
Years	1970-1999	1970-1999	1970-1999	1970-1999
N	2,361	2,074	2,074	1,866
Countries	127	123	123	115
PLL	-232.9	-223.8	-182.3	-169.0

NOTE: Robust standard errors clustered by country. Z-statistics in parenthesis.

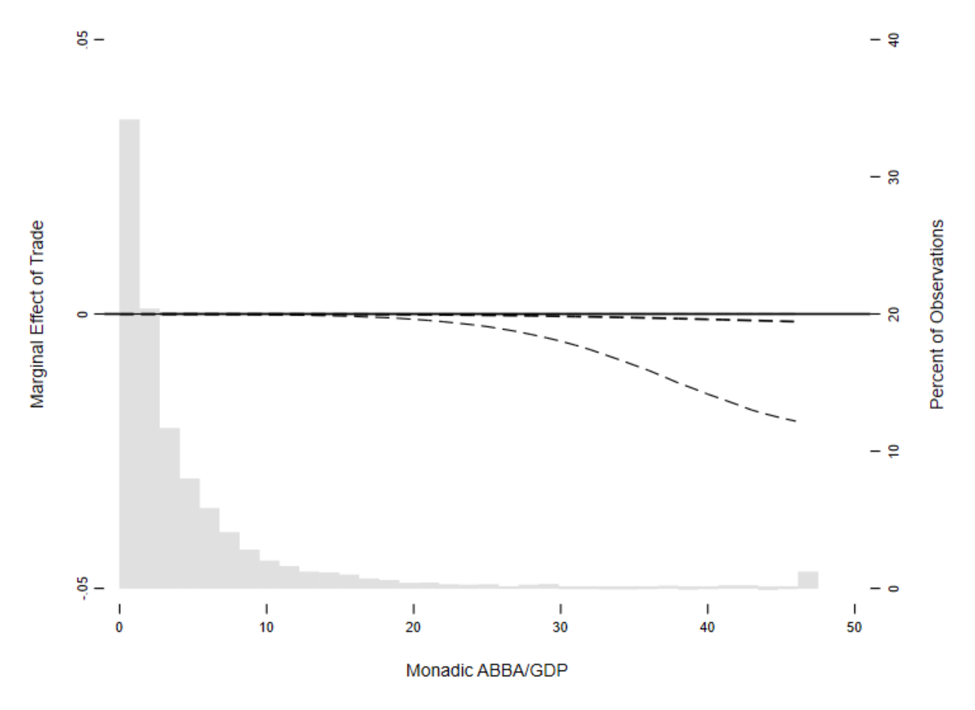
The remaining two models perform ABBA sensitivity checks. They clearly indicate that measurement problems matter. The z-statistic of the trade variable drops substantially when the ABBA term is included (Model 3), and the relationship falls below conventional levels of statistical significance when, in Model 4, we exclude the country-years in the top decile of the monadic ABBA distribution (in this case, observations in which it exceeds a sizeable 18.1 percent of GDP).[91]

Figure 9 plots the interaction between the reported trade effect and the monadic ABBA term. There is little correlation between trade and the risk of civil war presence when and where measurement errors are reasonably low. The original negative relationship between trade and civil

[91] Similarly, the relationship between trade and conflict presence weakens in substantive and statistical significance as we move, in the subset of trade flows with two independent mirror records, from the sum of bilateral trade as recorded by the reporting economy to that of partner economies and, finally, the sum of quality-weighted averages of the two (Models 6 to 8 in appendix table A13).

war presence, therefore, may well be driven by a modest number of observations for which mirror discrepancies, captured by the ABBA factor, are very high. It is conceivable that forces triggering violence also spawn inaccurate statistics. In this sense, our results do not necessarily invalidate the original findings or the important theoretical argument informing the empirical work. Nonetheless, the replication highlights how questionable trade statistics may complicate statistical study of the relationship between trade and conflict.[92]

Figure 9. Effect of a one unit increase in trade/GDP on the probability of civil war presence at different levels ABBA



NOTE: Graph produced using the code of Berry et al.[93]. For better readability, the maximum for the ABBA term was fixed at its 99th percentile in underlying regressions. All other variables are set at median value. Dotted lines indicate 95 percent confidence interval.

[92] Cf. Schultz 2015.
[93] Berry, Golder, and Milton 2012.

4.2.2. Garrett and Mitchell (2001)

The study of civil wars tends to focus on jurisdictions with often limited statistical capacity. Other debates using monadic trade data concentrate on advanced industrial economies. One prominent strand links economic openness and welfare spending. To assess measurement problems in these setups, we reconstruct the main models of a widely cited study by Garrett and Mitchell.[94] The authors investigate how globalization affects welfare states; here we concentrate on their analysis linking trade to general social policy spending. Garrett and Mitchell find *general* trade openness to be associated with (substantively small but statistically significant) *decreases* in such spending, while growing trade inflows *from low-wage economies* were associated with *increases*.[95]

With a dataset provided by Busemeyer, we follow Garrett and Mitchell's research design as closely as possible.[96] We undertake a few modifications to be able to illustrate the effect of trade data quality: we focus only on trade (not also FDI and portfolio flows) in the post-1980 period of interest in the original studies.[97] We standardize low-wage imports by GDP rather than total imports in order to remove trade measurement problems from the denominator. Also, the exclusion of low-quality data points makes the dataset too unbalanced for the calculation of panel-clustered standard errors, so we employ robust standard errors clustered at the country level instead.[98]

[94] Garrett and Mitchell 2001.

[95] See also Burgoon 2001.

[96] Other scholars have critiqued Garrett and Mitchell's econometric approach, for instance because it focusses on within variation and tends to smooth over cross-sectional variation and relationships, and because it simultaneously includes a lagged dependent variable and country fixed effects in the main models. The re-analysis by Kittel and Winner (2005) suggests that the findings do not hold in dynamic first-difference error-correction regressions, their preferred model, although the negative relationship of total trade re-appears when one uses time-series extending into the 2000s (Busemeyer 2009). These are important specification issues. But we sidestep them here because they swing free from the question to what degree trade data problems affect the conclusions *within* Garrett and Mitchell's modeling choices.

[97] That makes our replication a conservative test of the original findings' robustness, given that FDI and portfolio investment data quality is generally worse than that of trade data. Damgaard and Elkjaer 2014.

[98] Note that in our setup this leads to marginally smaller standard errors, which works against a rebuttal of the original findings.

We have to isolate the mirror problem for trade with low wage countries to replicate Garrett and Mitchell's (2001) finding about such trade. We create separate ABBA terms for total trade volumes and imports from low-wage countries. They parallel the monadic ABBA terms above, but the low-wage measure is limited to imports from non-OECD and non-OPEC economies. Figure 10 illustrates the resulting two monadic ABBA terms for the United States graphically. The discrepancies to watch are the gaps between the two lower lines. In both cases they grow over the decades, and US figures typically outstrip those of its partner countries.

Figure 10. Illustration of ABBA factor and low-wage ABBA factor for the USA

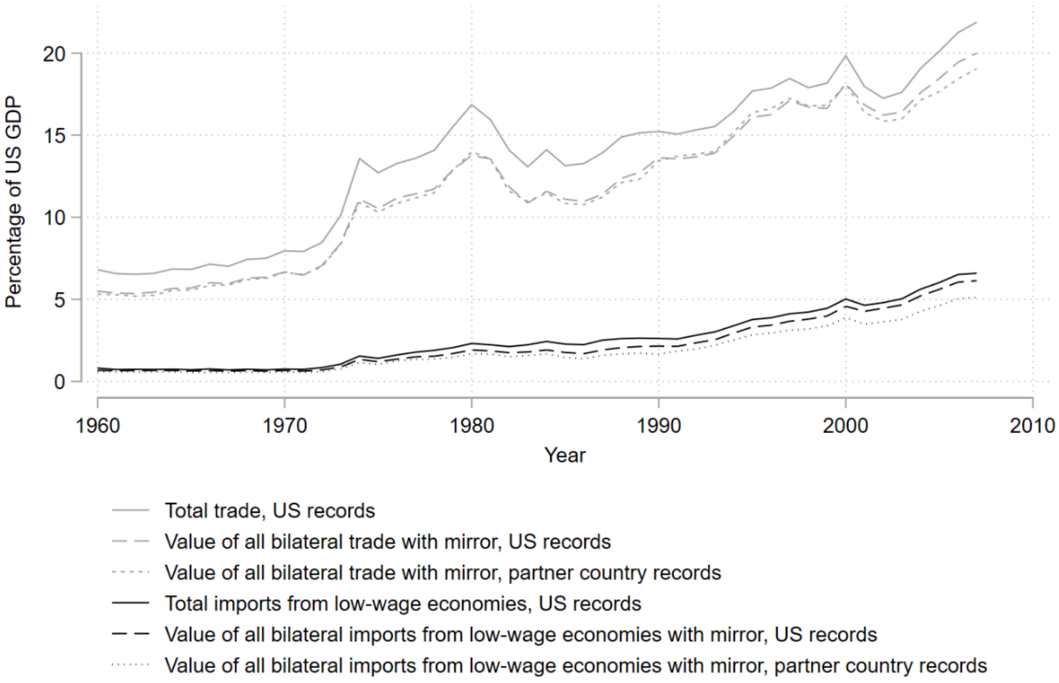


Table 7 summarizes our replications. We re-establish the original baseline (Model 1) and the baseline with merchandise (as opposed to total) trade (Model 2);[99] we include the monadic

[99] The included control variables are measures of deindustrialization, unemployment, GDP per capita, GDP growth, the dependency ratio, the share of cabinet positions held by left parties and the share held by Christian Democratic parties (full results are shown in Table A14 in the appendix).

ABBA term as a control (Model 3), and re-run the baseline in a restricted model that excludes the decile of country-years with highest ABBA terms (in the OECD sample these are countries with monadic ABBA terms exceeding a sizable 6.1 percent of GDP). Finally, we interact the trade variable with the underlying ABBA terms.

Table 7. Replication of Garrett and Mitchell (2001)

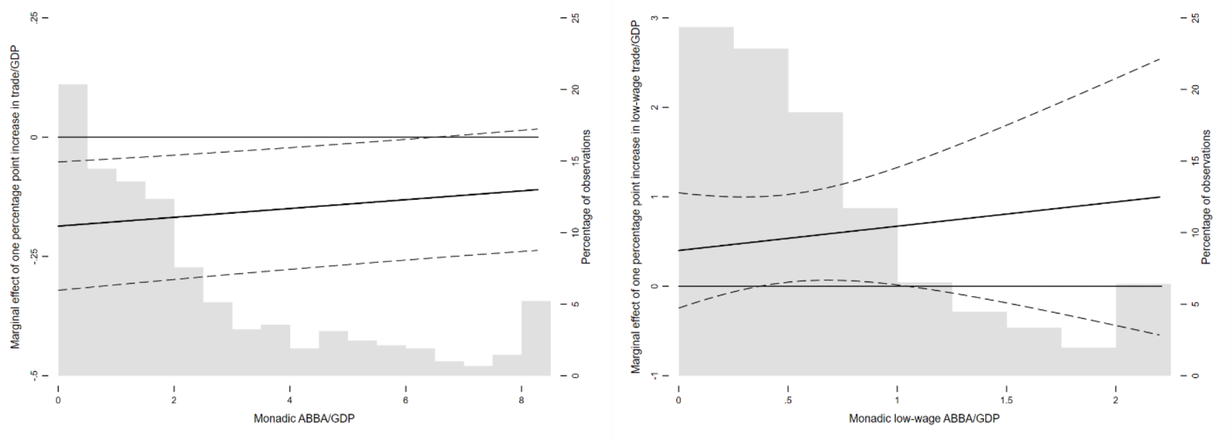
DV: Total spending/GDP	(1)	(2)	(3)	(4)
Model	*Baseline*	*Baseline merchandise trade*	*Monadic ABBA as control*	*Censoring ABBA top decile*
Total trade/GDP (t-1)	-0.10 (-2.69)			
Total merchandise trade/GDP (t-1)		-0.15 (-2.18)	-0.15 (-2.17)	-0.20 (-2.82)
ABBA factor (t-1)			0.01 (0.10)	
Low-wage imports/GDP (t-1)		0.42 (1.83)	0.45 (1.37)	0.39 (1.63)
Low-wage ABBA factor (t-1)			-0.14 (-0.40)	
Control variables included?	Yes	Yes	Yes	Yes
Country-fixed effects?	Yes	Yes	Yes	Yes
Year-fixed effects?	Yes	Yes	Yes	Yes
Years	1981-94	1981-94	1981-94	1981-94
Countries	21	19	19	19
N	258	240	240	219
R^2	0.98	0.98	0.98	0.98

NOTE: Robust standard errors clustered by country; t-statistic in parentheses.

Remarkably, the ABBA robustness checks pull in different directions for the two trade measures: the negative relationship between total trade and public spending waxes; the positive effect of low-wage imports wanes, with the coefficient losing statistical significance at the ten percent level.

An interaction between the relevant ABBA terms and the measures of trade (left-hand panel of Figure 11) or low-wage trade (right-hand panel) clarifies this pattern. For the total trade variable (left panel) the negative relationship is significantly negative when the potential for measurement error due to mirror discrepancies is small. For low-wage imports (right panel), in contrast, the positive relationship is strongest for countries with medium-low ABBA factors; it is small in substantive terms for the highest data-quality observations. In the latter case, then, low-quality data points seem to lead to an upward bias in the original estimates of the relationship.

Figure 11. Marginal effect of total trade (left) and imports from low-wage economies (right) on social spending at different values of data quality



NOTE: Graph produced using the code of Berry et al.[100]. For better readability, the maximum values for the ABBA terms are fixed at their 95th percentile in underlying regressions. Dotted lines indicate 95 percent confidence interval.

Taken together, measurement errors likely alter the modeled relationships between trade and welfare spending. Our replications suggest that the original studies may have underestimated

[100] Berry, Golder, and Milton 2012.

the negative relationship between trade openness and public spending. At the same time, they cast doubt on the robustness of the positive effect of low-wage imports. These patterns are quite one-sided and hence go against the more basic attenuation biases suggested by earlier replication studies focused on estimators and error-correction (e.g. Kittel and Winner 2005).

5. Implications and conclusion

IR and IPE scholarship has hitherto ignored or downplayed mirror discrepancies in trade data. Our analysis of such discrepancies yields three analytical insights and two recommendations. First, we have detailed the scale of the mirror problem. We have quantified the gaps between any two countries' estimates about their bilateral trade to construct ABBA terms as proxies for error in the data. When we zoom in on particular cases, such as US-Mexican trade, and when we conduct large-n analyses of such ABBA terms, we find significant, and sometimes massive, uncertainty in trade data. Unreflective choice for one or the other trade measure is problematic in and of itself: no particular trade measure is consistently and obviously superior to all others, regardless of whether we study bilateral trade or a country's trade with the rest of the world.

Second, we have investigated the origins of mirror discrepancies. If we could systematically account for discrepancies, we might control for their sources in statistical analyses, too. If, in contrast, they were completely random, we could treat them as noise in the data. Neither approach, alas, fits our overview of the data. Case studies of specific dyads and qualitative evidence suggest that the discrepancies are systematic and driven by particular features of the global economy, for example trade hubs, secrecy jurisdictions, and hard-to-track trade within multinational corporations. Yet because these factors confound trade data simultaneously, we cannot fully disentangle their contribution to specific discrepancies. Biases in trade data are therefore hard to eradicate. Statisticians try, for example through bilateral reconciliation exercises or the OECD's Trade-in-Value-Added (TiVA) database. But given the resource- and time-intensity of this work, the speed of change in the global economy, and the fundamental statistical capacity defects in many places, these initiatives clearly offer no short-term panacea.[101]

Third, mirror discrepancies may challenge scholarly knowledge of trade's origins and implications. Heeding mirror discrepancies affects what we think we know about trade and international conflict and political economy: it can strengthen or altogether wash-out the statistical significance of previous results; it can substantially change the magnitude of estimated effects; and in some cases it can even reverse their direction.

Two recommendations follow: first, future scholarship should explicitly take the mirror problem into account. This is easy for individual bilateral axes. Discussions of, say, Chinese-American trade should consider both sides of the mirror data and try to understand what drives data discrepancies. Matters are less straightforward for the large-n scholarship we have explored, but our replications suggest several easily implementable approaches to gauge the robustness of inference.[102] They include sample decomposition and re-measuring trade relationships through a "mirror substitution check" and the use of weighted mirror averages. We can also include control-variables that proxy discrepancies, such as the ABBA terms. Visualizing interactive relationships between trade and data quality is relatively straight-forward, and it reveals when and where mirror discrepancies affect statistical inference. To facilitate such robustness checks, this article is accompanied by publicly available datasets with the dyadic and monadic ABBA measures derived from IMF DOTS for a large swath of countries from 1950-2014 that will be periodically updated as new data becomes available. Even though we have limited our examples to IR and IPE

[101] Statisticians themselves remain sceptical about the ultimate promise of such endeavours. Mügge and Linsi 2020.
[102] Cf. Neumayer and Plümper 2017.

scholarship, both the problems we signal and the fixes we suggest are also relevant to work in international economics and business, fields that also frequently uses trade data.

The second and more broad-ranging recommendation is that IR and IPE scholarship take more seriously measurement problems in political economy more generally. Of the different quantities tracked in Balance of Payments-data—also including services trade, foreign direct investment and portfolio investment—merchandise trade is arguably the most reliable.[103] If things are as problematic for merchandise trade as our analyses show them to be, we should expect them to be worse for other facets of international economic relations. The time seems ripe to make critical discussion of data quality and measurement problems in official statistics a standard building block of academic training.

Data problems are not limited to cross-border exchange. For this article we have leveraged mirror discrepancies in international economic statistics. Yet statisticians and scholars have equally raised serious questions about measurement problems for domestic macroeconomic variables, as well: growth,[104] inflation,[105] public debt,[106] unemployment,[107] productivity,[108] and so on. Many scholars using these data ignore or smooth-over basic measurement concerns. It obviously goes beyond the confines of this article to tackle such concerns in detail. But given the systematic biases in official data, we should engage with measurement problems in key economic aggregates constructively and pro-actively so as to improve our inferences. At stake is the basic quality of what we know and argue about international and comparative political economy.

[103] Lipsey 2009; Damgaard and Elkjaer 2014; Linsi and Mügge 2019.
[104] Brynjolfsson, Eggers, and Gannamaneni 2018; UNECE, Eurostat, and OECD 2011.
[105] Boskin et al. 1998; Mackie and Schultze 2002.
[106] Bloch and Fall 2015.
[107] Hoskyns and Rai 2007.
[108] Guvenen et al. 2017; Brynjolfsson, Rock, and Syverson 2017.

References

Alves, John. 1967. Progress Towards Uniformity in Balance of Payments Presentation. In *Research and Statistics Department, Assistant Director Arie Bouter Files, Box #1 File #8*. Washington, D.C.: IMF Archives.

Barbieri, Katherine, and Omar Keshk. 2011. Too Many Assumptions, Not Enough Data. *Conflict Management and Peace Science* 28 (2): 168–172.

Barbieri, Katherine, Omar M G Keshk, and Brian M Pollins. 2009. Trading Data: Evaluating our Assumptions and Coding Rules. *Conflict Management and Peace Science* 26 (5). SAGE Publications Ltd: 471–491. Available at <https://doi.org/10.1177/0738894209343887>.

Barbieri, Katherine, and Rafael Reuveny. 2005. Economic Globalization and Civil War. *Journal of Politics* 67 (4): 1228–1247.

Berry, William D, Matt Golder, and Daniel Milton. 2012. Improving Tests of Theories Positing Interaction. *The Journal of Politics* 74 (3). The University of Chicago Press: 653–671. Available at <https://doi.org/10.1017/S0022381612000199>.

Bhagwati, Jagdish. 1967. Fiscal policies, the faking of foreign trade declarations, and the balance of payments. *Bulletin of the Oxford University Institute of Economics & Statistics* 29 (1). John Wiley & Sons, Ltd: 61–77. Available at <https://doi.org/10.1111/j.1468-0084.1967.mp29001004.x>.

Bhagwati, Jagdish. 1964. On the underinvoicing of imports. *Bulletin of the Oxford University Institute of Economics & Statistics* 27 (4). John Wiley & Sons, Ltd: 389–397. Available at <https://doi.org/10.1111/j.1468-0084.1964.mp27004007.x>.

Bloch, Debra, and Falilou Fall. 2015. *Government Debt Indicators. Understanding the Data*. OECD Economics Department Working Papers. Paris.

Boehmer, Charles R., Bernadette Jungblut, and Richard J. Stoll. 2011. Tradeoffs in Trade Data: Do Our Assumptions Affect Our Results? *Conflict Management and Peace Science* 28 (2): 145–167.

Boskin, Michael, Ellen Dulberger, Robert Gordon, Zvi Griliches, and Dale Jorgenson. 1998. Consumer Prices, the Consumer Price Index, and the Cost of Living. *Journal of Economic Perspectives* 12 (1): 3–26.

Braml, Martin T., and Gabriel J. Felbermayr. 2019. *The EU Self-Surplus Puzzle: An Indication of VAT Fraud?* CESifo Working Paper. Munich. Available at <https://www.ifo.de/en/node/51244>.

Broome, André, and Joel Quirk. 2015. The politics of numbers: the normative agendas of global benchmarking. *Review of International Studies* 41 (05): 813–818. Available at <http://www.journals.cambridge.org/abstract_S0260210515000339>.

Brynjolfsson, Erik, Felix Eggers, and Avinash Gannamaneni. 2018. Measuring Welfare with Massive Online Choice Experiments: a Brief Introduction. *AEA Papers and Proceedings* 108 (May): 473–476.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2017. *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*. NBER Working Paper Series. Cambridge MA. Available at <https://www.nber.org/papers/w24001>.

Burgoon, Brian. 2001. Globalization and Welfare Compensation: Disentangling the Ties That Bind. *International Organization* 55 (3): 509–551.

Busemeyer, Marius R. 2009. From myth to reality: Globalisation and public spending in OECD countries revisited. *European Journal of Political Research* 48 (4): 455–482.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman & Hall.

Damgaard, Jannick, and Thomas Elkjaer. 2014. Foreign Direct Investment and the External Wealth of Nations: How Important is Valuation? *Review of Income and Wealth* 60 (2): 245–260.

Ely, Edward J. 1961. Variations Between U.S. and Its Trading Partner Import and Export Statistics. *The American Statistician* 15 (2): 23–26.

Eurostat. 2016. *User guide on European statistics on international trade in goods*. Luxembourg: Publications Office of the European Union. Available at <https://ec.europa.eu/eurostat/documents/3859598/7679615/KS-GQ-16-009-EN-N.pdf/073b853a-a4f4-4c55-aaba-162671544c78>.

Fearon, James D., and David D. Laitin. 2003. Ethnicity, Insurgency, and Civil War. *American Political Science Review* 97 (1): 75–90.

Fortanier, Fabienne, and Katia Sarrazin. 2016. *Balanced International Merchandise Trade Data: Version 1*. Paris. Available at <https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=STD/CSSP/WPTGS%282016)18

&docLanguage=En>.

Garber, Molly E., Ted Peck, and Kristy L. Howell. 2018. Understanding Asymmetries Between BEA's and Partner Countries' Trade Statistics. *Survey of Current Business: The Journal of the U.S. Bureau of Economic Analysis* 98 (2). Available at <https://apps.bea.gov/scb/2018/02-february/pdf/Asymmetries-in-Bilateral-Trade-Statistics_SCB-February-2018.pdf>.

Garrett, Geoffrey, and Deborah Mitchell. 2001. Globalization, government spending and taxation in the OECD. *European Journal of Political Research* 39 (2): 145–177.

Gaspareniene, Ligita, Rita Remeikiene, and Friedrich Georg Schneider. 2015. The factors of digital shadow consumption. *Intellectual Economics* 9: 109–119. Available at <https://www3.mruni.eu/ojs/intellectual-economics/article/view/4413>.

Gaulier, Guillaume, and Soledad Zignago. 2010. *BACI: International Trade Database at the Product-Level: The 1994-2007 Version*. CEPII Working Paper. Available at <http://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?NoDoc=2726>.

Gehlhar, Mark J. 1996. *Reconciling Bilateral Trade Data for Use in GTAP*. GTAP Technical Paper No. 10. Available at <https://www.gtap.agecon.purdue.edu/resources/download/38.pdf>.

Gleditsch, Kristian Skrede. 2010. On Ingoring Missing Data and the Robustness of Trade and Conflict Results: A Reply to Barbieri, Keshk and Pollins. *Conflict Management and Peace Science2* 27 (2): 153–157.

Goldstein, Judith L., Douglas Rivers, and Michael Tomz. 2007. Institutions in International Relations: Understanding the Effeccts of the GATT and the WTO on World Trade. *International Organization* 61 (1): 37–67.

Guvenen, Fatih, Raymond Mataloni, Dylan Rassier, and Kim Ruhl. 2017. *Offshore Profit Shifting and Domestic Productivity Measurement*. NBER working paper series. Cambridge MA. Available at <https://www.nber.org/papers/w23324>.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2011. Democracy and Transparency. *Journal of Politics* 73 (4): 1191–1205.

Hoskyns, Catherine, and Shirin Rai. 2007. Recasting the Global Political Economy: Counting Women's Unpaid Work. *New Political Economy* 12 (3): 297–317.

International Monetary Fund. 1993. *A Guide to Direction of Trade Statistics*. Washington, D.C.

International Monetary Fund. 1987. *Report on the World Current Account Discrepancy*. Washington, D.C.

Javorsek, Marko. 2016. *Asymmetries in International Merchandise Trade Statistics: A case study of selected countries in Asia-Pacific*. Statistics Division Working Paper Series. Available at <https://ec.europa.eu/eurostat/documents/7828051/8076585/Asymmetries__trade_goods.pdf>.

Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Ithaca: Cornell University Press.

Kastner, Scott L. 2016. Buying Influence? Assessing the Political Effects of China's International Trade. *Journal of Conflict Resolution* 60 (6): 980–1007.

Kelley, Judith G., and Beth A. Simmons. 2019. Introduction: The Power of Global Performance Indicators. *International Organization* 73 (3): 491–510.

Kerner, Andrew. 2014. What We Talk About When We Talk About Foreign Direct Investment. *International Studies Quarterly* 58 (4): 804–815.

Kim, In Song, Steven Liao, and Kosuke Imai. 2020. Measuring Trade Profile with Granular Product-Level Data. *American Journal of Political Science* 64 (1). John Wiley & Sons, Ltd: 102–117. Available at <https://doi.org/10.1111/ajps.12473>.

Linsi, Lukas, and Daniel Mügge. 2019. Globalization and the growing defects of international economic statistics. *Review of International Political Economy* 26 (3): 361–383.

Lipsey, Robert E. 2009. Measuring International Trade in Services. In *International Trade in Services and Intangibles in the Era of Globalization*, edited by Marshall Reinsdorf and Matthew J. Slaughter, 27–70. Chicago: University of Chicago Press. Available at <http://www.nber.org/chapters/c11605.pdf>.

Mackie, Christopher, and Charles Schultze. 2002. *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes*. Washington DC: National Academies Press.

Markhonko, Vladimir. 2014. Asymmetries in official international trade statistics and analysis of globalization. In *International Conference on the Measurement of International Trade and Economic Globalization*, 1–17. Aguascalientes, Mexico. Available at <https://unstats.un.org/unsd/trade/events/2014/mexico/Asymmetries in official ITS and analysis of globalization - V Markhonko - 18 Sep 2014.pdf>.

Miao, Guannan, and Fabienne Fortanier. 2017. Estimating Transport and Insurance Costs of International Trade. *OECD Statistics Working Papers* (4). Available at <http://dx.doi.org/10.1787/18152031>.

Morgenstern, Oskar. 1963. *On the accuracy of economic observations*. Second edi. Princeton, N.J.: Princeton

University Press.

Mügge, Daniel, and Lukas Linsi. 2020. The national accounting paradox: how statistical norms corrode international economic data. *European Journal of International Relations*. SAGE Publications Ltd: 1354066120936339. Available at <https://doi.org/10.1177/1354066120936339>.

Naya, Seiji, and Theodore Morgan. 1969. The Accuracy of International Trade Data: The Case of Southeast Asian Countries. *Journal of the American Statistical Association* 64 (326). Taylor & Francis: 452–467. Available at <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1969.10500987>.

Neumayer, Eric, and Thomas Plümper. 2017. *Robustness Tests for Quantitative Research*. Cambridge: Cambridge University Press.

Office for National Statistics. 2020. *Asymmetries in trade data: updating analysis of UK bilateral trade data*. London. Available at <https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/articles/asymmetriesintradedatadivingdeeperintoukbilateraltradedata/updatinganalysisofukbilateraltradedata>.

Osgood, Iain. 2017. The Breakdown of Industrial Opposition to Trade: Firms, Product Variety, and Reciprocal Liberalization. *World Politics* 69 (1). Cambridge University Press: 184–231.

Rose, Andrew K. 2004. Do We Really Know That the WTO Increases Trade? *American Economic Review* 94 (1): 98–114.

Schultz, Kenneth A. 2015. Borders, Conflict, and Trade. *Annual Review of Political Science* 18: 125–145.

Smith, John S. 1966. Asymmetries and Errors in Reported Balance of Payments Satistics. In *Research and Statistics Department, Assistant Director Arie Bouter Files, Box #1 File #8*. Washington, D.C.: IMF Archives.

UNECE, Eurostat, and OECD. 2011. *The Impact of Globalization on National Accounts*. New York and Geneva: United Nations.

United Nations Statistics Division. 2011. *International Merchandise Trade Statistics: Concepts and Definitions 2010*. New York. Available at <https://unstats.un.org/unsd/trade/eg-imts/IMTS 2010 (English).pdf>.

Yeats, Alexander J. 1990. On the Accuracy of Economic Observations: Do Sub-Saharan Trade Statistics Mean Anything? *The World Bank Economic Review* 4 (2): 135–156.

Yeats, Alexander J. 1978. On the accuracy of partner country trade statistics. *Oxford Bulletin of Economics and Statistics* 40 (4). John Wiley & Sons, Ltd: 341–361. Available at <https://doi.org/10.1111/j.1468-0084.1978.mp40004004.x>.

Ylönen, Matti, and Teivo Teivanen. 2018. Politics of Intra-firm Trade: Corporate Price Planning and the Double Role of the Arm's Length Principle. *New Political Economy* 23 (4): 441–457.

1956. Verbatim Report of the International Monetary Fund Meeting of Fund Statistical Correspondents on Balance of Payments Discussions held at the Burgundry Room - Sheraton-Park Hotel, Washington D.C. on Thursday, September 27, 1956 at 2:30 pm. In *Research and Statistics Department, Assistant Director Arie Bouter Files, Box #1 File #1*. Washington, D.C.: IMF Archives.