The Problem with Trade Measurement in International Relations

Linsi, Lukas; Burgoon, Brian; Mügge, Daniel

Published in:
International Studies Quarterly

DOI:
10.1093/isq/sqad020

International Studies Quarterly (2023) **67**, sqad020

The Problem with Trade Measurement in International Relations

Lukas Linsi
University of Groningen, The Netherlands

and

Brian Burgoon and Daniel K. Mügge
University of Amsterdam, The Netherlands

Trade statistics are widely used in studies and policymaking focused on economic interdependence. Yet, researchers in International Relations (IR) have largely disregarded half the data available to study trade. Bilateral trade flows are usually recorded twice: by the sending economy as an export and by the receiving one as an import. These two values should match, but discrepancies between them tend to be large and pervasive. Most studies ignore this issue, which we label the "mirror problem" for short, by using only one entry. However, it is not self-evident which one is consistently most accurate. Hence, IR's reliance on error-prone trade statistics may be distorting its study of economic interdependence. This article explores this problem in three steps: first, we quantify the mirror problem in trade data. Second, we investigate the origins of the mirror problem, using statistical analyses, archival records, and interviews with statistical experts. Third, we illustrate the implications of the mirror problem through replications covering diverse topics in IR. We find that accounting for the mirror problem can variably strengthen, undermine, or overturn conclusions of such analyses. The findings underscore the severity of measurement problems in IR and suggest particular ways to address those problems.

Las estadísticas comerciales se utilizan con mucha frecuencia en los estudios y en la elaboración de políticas centrados en la interdependencia económica. Sin embargo, los investigadores en materia de Relaciones Internacionales (RRII) han ignorado en gran medida la mitad de los datos disponibles para estudiar el comercio. Los flujos comerciales bilaterales suelen registrarse dos veces: por parte de la economía emisora como exportación y por parte de la economía receptora como importación. Estos dos valores deberían coincidir, pero las discrepancias entre ellos suelen ser grandes y generalizadas. La mayoría de los estudios ignoran este problema, que denominamos «problema del espejo», utilizando solo una de estas entradas. Pero no resulta evidente cuál de ellos es el más preciso. Por lo tanto, la dependencia por parte de las RRII de estadísticas comerciales propensas a errores puede estar distorsionando su estudio de la interdependencia económica. Este artículo analiza este problema en tres pasos: En primer lugar, cuantificamos el problema del espejo en los datos comerciales. En segundo lugar, investigamos los orígenes del problema del espejo, utilizando análisis estadísticos, registros de archivo y entrevistas con expertos en estadística. En tercer lugar, ilustramos las implicaciones del problema del espejo mediante repeticiones que abarcan diversos temas de las RRII. Comprobamos que tener en cuenta el problema del espejo puede reforzar, socavar o anular las conclusiones de dichos análisis. Los resultados subrayan la gravedad de los problemas de medición en las RRII y sugieren formas concretas de abordarlos.

Les études et l'élaboration de politiques centrées sur l'interdépendance économique ont largement recours aux statistiques commerciales. Pourtant les chercheurs en relations internationales (RI) ont jusqu'ici largement ignoré la moitié des données disponibles pour l'étude du commerce. En général, les flux commerciaux bilatéraux sont enregistrés deux fois : une fois par l'économie émettrice en tant qu'exportation, et une autre par la destinatrice en tant qu'importation. Ces deux valeurs devraient correspondre, mais les différences ont tendance à être importantes et répandues. La plupart des études omettent cette problématique, que nous appelons « le problème du miroir » pour faire court, en n'utilisant qu'une seule entrée. Or, l'entrée à la précision le plus constante n'est pas toujours évidente. Ainsi, la dépendance des RI sur des statistiques commerciales où les erreurs sont fréquentes est susceptible de fausser son étude de l'interdépendance économique. Le présent article étudie ce problème en trois étapes. D'abord, nous quantifions le problème du miroir en données commerciales. Puis, nous nous intéressons aux origines du problème du miroir, à l'aide d'analyses statistiques, d'archives et d'entretiens avec des experts en statistiques. Enfin, nous illustrons les implications du problème du miroir à l'aide de reproductions couvrant différents sujets des RI. Nous observons que la prise en compte du problème du miroir peut renforcer les conclusions de l'analyse, mais aussi leur nuire ou les contredire. Nos conclusions soulignent l'importance des problèmes de mesure dans les RI avant de proposer des façons spécifiques d'y répondre.

Lukas Linsi is an Assistant Professor of International Political Economy at the University of Groningen. His research focuses on multinational corporations, international trade, and the politics of economic measurement.

Brian Burgoon is a Professor of International and Comparative Political Economy at the Department of Political Science at the University of Amsterdam. His research focuses on political responses to economic openness.

Daniel Mügge is a Professor of Political Arithmetic at the University of Amsterdam. He has been principal investigator of the FickleFormulas project about macroeconomic indicators. His current work focuses on the regulation of artificial intelligence.

Authors' note: Earlier versions of the article have been presented at the IPES 2020 conference and workshops at the University of Amsterdam and University of Zurich. We thank participants for many useful suggestions. We also gratefully acknowledge detailed comments we received from the anonymous reviewers and *International Studies Quarterly* editors. Their feedback has significantly improved the article. Michael Tomz, Catherine Barbieri, Marius Busemeyer, and Scott Kastner made available the replication data used in this article. Javier Garcia Bernardo and Frank Takes helped us with the preparation of the IMF DOTS dataset. Jessica di Salvatore, Maximilian Fenner, Roel van Engelen, and Stefan Sliwa Ruiz provided valuable research assistance. The research was generously supported financially by an Early Postdoc Mobility Grant from the Swiss National Science Foundation (grant P2SKP1_168289), the ERC Starting Grant FICKLEFORMS (grant #637883), and the NWO Vidi project 016.145.395. All errors remain our own. Replication files for the analyses undertaken in this article are available on the ISQ Dataverse, at https://dataverse.harvard.edu/dataverse/isq. Trade datasets that enable researchers to implement the recommendations made in this article for their own research are made available on the dedicated "Mirror Trade" dataverse page: https://dataverse.harvard.edu/dataverse/mirrortrade/.

Linsi, Lukas et al. (2023) The Problem with Trade Measurement in International Relations. *International Studies Quarterly*, https://doi.org/10.1093/isq/sqad020

Introduction

Trade statistics stand central in research on global economic governance and international relations (IR). 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 IR scholars, import and export figures are the most common measures of economic interdependence, crucial to understanding the character, origins, and implications of globalization.

Research designs to study the origins and consequences of trade have become ever more advanced, and extensively debated. While such debate has focused mainly on causal identification, scholars and policymakers have mostly disregarded basic defects of trade data itself. Some IR scholars have scrutinized the operationalization of trade interdependence (e.g., Gartzke and Li 2003; Gray and Potter 2012); others have questioned measures of complex interactions such as foreign direct investment (Kerner 2014) or trade in services (Weymouth 2017, 938). The basics of trade measurement, however, have been treated as rather unproblematic. Yet, as acknowledged by the International Monetary Fund (IMF) or the Organization for Economic Cooperation and Development (OECD) (International Monetary Fund 1987; UNECE, Eurostat, and OECD 2011), data quality can be lacking for trade in goods, too.

The so-called mirror statistics evidence such measurement uncertainties. Trade flows are in principle recorded twice, once as exports by the sending economy and once as imports by the receiving one. The IMF's Direction of Trade Statistics (DOTS) database,[1] the most widely used resource for bilateral trade statistics in IR research, provides both figures. If they were very similar, "mirror statistics" would not raise significant questions. Yet, discrepancies in mirror statistics are large and persistent, even between countries with highly developed statistical systems. In our global sample, *on average* discrepancies are almost as large as trade volume estimates themselves.[2] Dyads with modest trade are important drivers of these huge differences. But also for dyads that trade a lot, the differences remain substantial: among the dyads in the top decile in trade volume (more than USD 92.5 million in trade per year according to importing-country records), the median discrepancy, relative to the average of two mirror flows, is still 23.1 percent (with a mean of 40.2 percent).

This "mirror problem," as we shall call it, reveals the substantial uncertainty in trade statistics, challenging analyses of the character, origins, or implications of trade (cf. Morgenstern 1963; International Monetary Fund 1987; Schultz 2015; Linsi and Mügge 2019). Trade statisticians and economists have long recognized mirror discrepancies (Ely 1961; Morgenstern 1963; Bhagwati 1964, 1967; Yeats 1978, 1990; Gaulier and Zignago 2010). Some have proposed statistical remedies, such as estimating mirror averages weighted by inferred reporter reliability in the BACI[3] or OECD Balanced International Merchandise Trade Statistics (BIMTS)[4] datasets. While we recognize and build on these efforts, their coverage remains too limited for many IR analyses—most of which ignore trade data defects altogether.[5] Using the most widely available trade statistics, often based on import values alone, most IR scholars implicitly trust those measures to be the "right" ones, discarding potentially valuable information from parallel export records. Yet, as we argue below, the assumption that estimates based on import records are more reliable frequently does not hold. Leveraging information from both sides of the mirror—rather than mechanically relying on only one side—can improve trade volume estimates and inferences drawn from them.

This paper explores the nature and origins of the mirror problem and its implications for research. 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 datasets of errors in common trade measures that we make publicly available with this paper.[6] These datasets reveal large and persistent discrepancies not confined to specific countries or regions of the world.

Second, we explore the sources of these measurement problems. Archival research and interviews with leading trade statisticians highlight the complexity of 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 our 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 research. We replicate five studies chosen to capture key varieties of IR topics and statistical setups:[7] the studies feature trade as both explanation and outcome, in both security and political economy issues, and in both dyadic and monadic settings. They include, in particular, studies of how economic globalization affects welfare states and of how multilateral institutions shape actual trade interdependence, and they include studies of links between trade and geopolitical alignments and military conflicts.

We find that the mirror problem matters a lot for IR findings. Accounting for measurement error can significantly strengthen or weaken statistical significance of results. It frequently changes the magnitude of estimated effects substantially, and in some cases reverses their direction. We identify easily implementable recommendations for using mirror trade information to improve the validity of trade-related findings. First, when analyzing bilateral trade between *two or a few countries*, we encourage IR scholars to compare import-record-based trade values with export-record-based ones whenever mirror records are available. In case of discrepancies, we recommend investigating their roots—which are often idiosyncratic and dyad-specific. For dyads with mirror records, we also recommend considering inferred-quality-weighted average values that can be more reliable than either side of the mirror alone. Second, for large-*n* studies, we recommend robustness checks focused on statistical discrepancies in trade flows. For trade flows in *large-n dyadic* datasets, we provide accompanying datasets

[1] Available here: http://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85 (Last accessed March 21, 2023).

[2] The median size of mirror discrepancies, normalized by the average of trade volumes reported by importers and exporters, is 79.6 percent (the mean 98.9 percent). Even if we drop all records in which either of the partner countries reports actual zero trade, the median is 47.2 percent and the mean 69.0 percent.

[3] Available here: http://www.cepii.fr/cepii/en/bdd_modele/presentation.asp?id=37 (Last accessed March 21, 2023).

[4] Available here: https://stats.oecd.org/Index.aspx?DataSetCode=BIMTS_CPA (Last accessed March 21, 2023).

[5] 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), and Schultz (2015).

[6] The original version accompanying this article covers the years 1950–2014. The original as well as an updated version covering the years 1948–2021 are available on the following link: https://dataverse.harvard.edu/dataverse/mirrortrade.

[7] These studies are Rose (2004), Barbieri and Reuveny (2005), Goldstein, Rivers, and Tomz (2007), and Garrett and Mitchell (2001), and Kastner (2016).

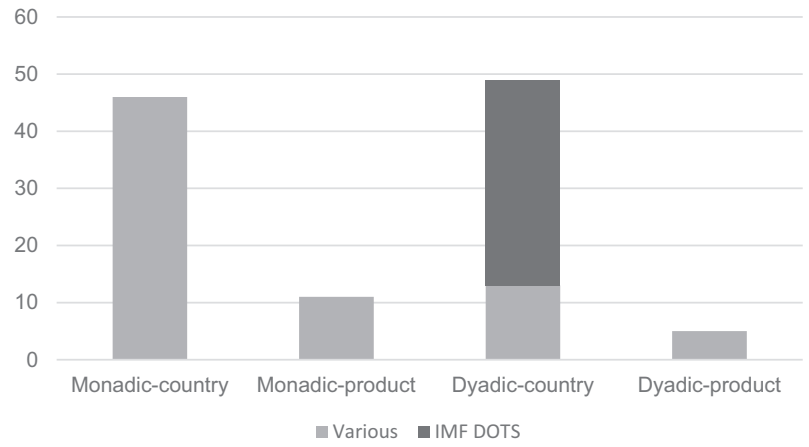

Figure 1. Trade data use in six leading journals, 2013–2017.
Source: Data collected by authors from journal homepages (details in text).

that allow IR researchers to compare results using import- with export-records-based trade volumes, as well as with an inferred-quality-weighted average of the two, in the large sample of dyads with mirror records. For large-*n* monadic datasets, we also identify ways to gauge and address bilateral mirror trade discrepancies, for which we generate and make available another accompanying dataset of trade discrepancies and inferred trade data quality for country-years. Including monadic trade error terms in regression models and plotting the interaction of trade variables with measures of inferred trade quality, as our replication exercises illustrate, clarify the direction and extent to which measurement problems bias baseline results in monadic studies.

Our message is also, however, that these recommendations are no panacea for the mirror problem. The robustness and sensitivity checks we detail below focus on the subsample of observations for which both trading partners publish independently estimated trade volume figures.[8] Since they represent the subsample of dyads for countries with relatively developed statistical apparatuses (cf. online appendix C), they will signal "lower end" indications of how measurement issues affect findings. Nonetheless, our recommendations can strengthen trade-related findings in IR by reducing "uncertainty about measurement uncertainty" and how it affects statistical analyses.

The Use of Trade Statistics in IR Research

Cross-border commerce stands central in IR research. We reviewed all articles between 2013 and 2017 in *International Organization*, *International Studies Quarterly*, *World Politics*, *Journal of Conflict Resolution*, *Journal of Politics*, and *European Journal of Political Research*. One hundred and eight articles use trade data, more than one in fifteen, almost all employing country-level data. Trade flows appear in four primary modes of analysis: monadic country (total imports/exports of a country); monadic product (imports/exports of goods in specific product categories); dyadic country (total flows among country pairs); and dyadic product (bilateral flows in specific product categories). Of the 108 studies, 49 used dyadic-country data and 46 monadic-country data. Product-level trade data remain rare in IR (cf. Kim, Liao, and Imai 2020), with eleven using monadic-product and five dyadic-product data (see figure 1).[9]

These studies rely on well-known data-gathering bodies. 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 United Nations [UN] sources); the remainder comes from US government and other sources (e.g., Eurostat), or is unspecified (15 percent). Monadic-product level studies draw primarily on the World Integrated Trade Solution (WITS) database. 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 based upon IMF DOTS). The remainder comes from UN, National Bureau of Economic Research, and sundry national or regional sources. On the whole, the reviewed studies take trade–data quality for granted, with little critical discussion or analysis beyond the occasional disclaimer.

Yet already in 1950, Oskar Morgenstern noted in *On the Accuracy of Economic Observations* 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" (Morgenstern 1963 [1950], 180). Bhagwati analyzed how over- or under-invoicing of trade biased balance of payments (BOP) data (Bhagwati 1964, 1967). Other scholars have lamented "often considerable" (Yeats 1978, 354) discrepancies in bilateral trade records (Naya and Morgan 1969; Yeats 1990; Braml and Felbermayr 2019). Statistical agencies and international organizations have also highlighted the mirror problem for years (International Monetary Fund 1987; International Monetary Fund 1993; Javorsek 2016; Garber, Peck, and Howell 2018; Office for National Statistics 2020), although the problem remains unresolved (Schultz 2015; Linsi and Mügge 2019).

Mirror Discrepancies and Their Uncertain Origins

How substantial are mirror discrepancies, and what might explain them? For a first impression, figure 2 visualizes the United States' merchandise trade deficit with Mexico—a politically highly salient figure. US statistics show it rising

[8] The IMF DOTS database with which we work includes information for 1,344,648 unidirectional trade flows, of which 518,517 have an independently recorded mirror flow. The latter correspond to 38.6 percent of all observations and cover 78.0 percent of total trade.

[9] Some articles use more than one type of trade data. Our summary excludes one study using firm-level data.

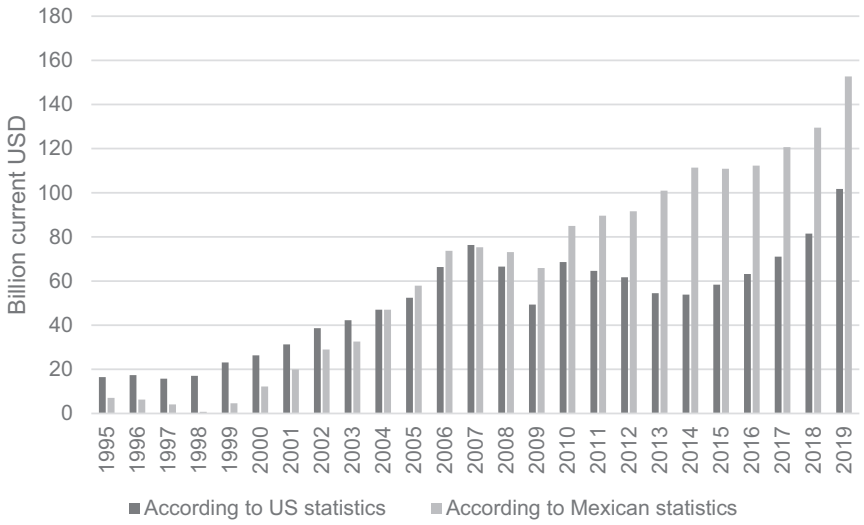

Figure 2. The size of the US merchandise trade deficit with Mexico, 1995–2019.
Source: Own calculations based on IMF DOTS database.

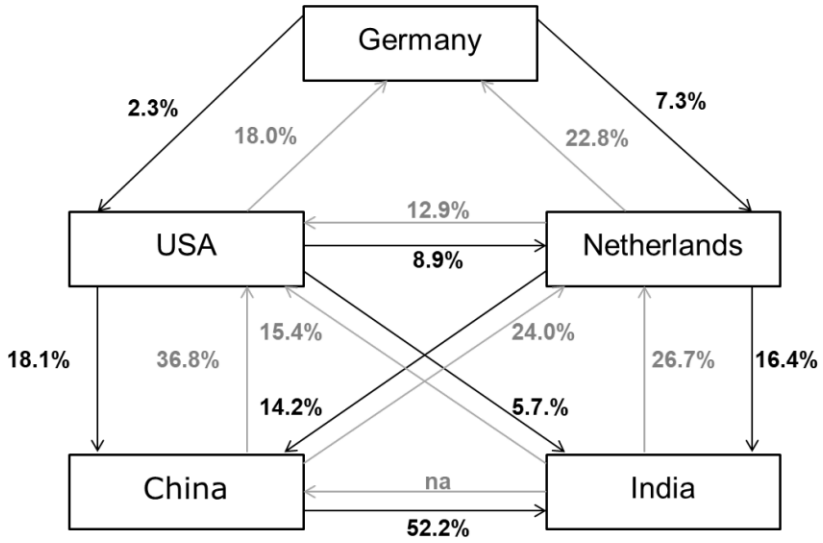

Figure 3. Mirror discrepancies as share of import value, 1995–2014 period averages.
Source: Own calculations based on IMF DOTS.

sharply from 1995 to 2007 and stabilizing afterward, although rising again after 2015. In Mexican data, the upward trend continues throughout the period, exceeding American ones by more than 50 percent after 2013. (The underlying calculations are detailed in online appendix table A1.)

Such discrepancies are not limited to the US–Mexico trade, as figure 3 illustrates. It depicts trade between the United States, Germany, the Netherlands, China, and India. The percentages indicate the discrepancy as a share of the total value of recorded imports, averaged over 20 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. This pattern shows up for larger country samples and time periods–with discrepancies being as or more substantial in more recent years than earlier periods, and among advanced industrialized as well as developing economies (see online appendix A). The message is clear: discrepancies are large and pervasive.

Mirror Discrepancies beyond Snapshots: ABBA Terms

To explore the underlying causes and consequences of discrepancies systematically, we need standardized measures. The measures we propose, for both dyad-years and country-years, gauge differences between what country A reports sending to country B and what B reports receiving from A (and vice versa)—"ABBA terms" for short. ABBA terms need different operationalizations for dyadic and monadic data. The *dyadic ABBA terms* can be defined 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, that generates two ABBA terms, one for each direction of trade. This initial definition is deliberately simple so that ABBA terms

can be used flexibly and adapted to the analytical context, for example, normalizing them by a common denominator.

Uncertain Origins of Mirror Discrepancies

What underlies the described discrepancies? And to what degree are they distributed nonrandomly, implying that we are not only dealing with poor data, but with systematically skewed images of global trade?

Potential explanations for discrepancies abound. First, "cost of insurance and freight" (c.i.f.) is included in import prices but not in the price of exports, which are loaded "free on board" (f.o.b.), although falling trade costs have shrunk such c.i.f.–f.o.b. differences over time (Miao and Fortanier 2017). Second, limited statistical capacities may drive mirror discrepancies (cf. Jerven 2013). Some countries have better-resourced data collection systems than others, and economic crises or wars can undermine data collection (Schultz 2015), exacerbating data disagreements. Third, accounting-technical glitches or cross-country differences in statistical practices can yield discrepancies (International Monetary Fund 1993). Fourth, trading entities face incentives to misreport the value of shipped goods. For instance, high tariff rates encourage under-invoicing of imports (Bhagwati 1964) and export subsidies the over-invoicing of exports (Bhagwati 1967). Over-invoiced imports can be used to circumvent capital controls (Yeats 1990), while European Union (EU) common market rules encourage over-invoicing exports to evade value-added tax payments (Braml and Felbermayr 2019). Fifth and finally, globalizing production is a key driver of discrepancies (UNECE, Eurostat, and OECD 2011; Linsi and Mügge 2019). International trading activities have become more complex due to multinational firms' growing reliance on global value chains (Baccini, Dür, and Elsig 2018; Kim et al. 2019), and often involve intermediate goods and merchanting that cause conflicting attributions between source and destination countries.[10]

It is difficult to attribute observed discrepancies to these drivers. Some are hard to observe or proxy, and proxies can pull in different directions. For instance, advanced economies can have high statistical capacity, but are also deeply integrated into complex global value chains that cloud estimates of trade flows. These ambiguities may be distortions that vary systematically. But because they distort data *simultaneously*, disentangling them is challenging—something noted by Yeats (1990, 136–37) and corroborated in interviews with OECD, World Trade Organization (WTO), and IMF statisticians.[11]

Furthermore, idiosyncrasies also drive discrepancies. Statistical reconciliation exercises reveal their roots to be diverse and specific to particular dyads. For instance, differential treatment of some intermediate components shipped between the United States and maquiladoras owned by US companies partly drove US–Mexican trade discrepancies before 2007 (figure 2), but the inverted discrepancies from 2007 onward remain less well understood.[12] Notable discrepancies in UK–Swiss bilateral goods trade were driven by differential classifications of trade in nonmonetary gold, and changes in ownership without any physical trade occurring (Office for National Statistics 2017). Reconciliation exercises of the German Statistical Office with EU partner institutions highlighted intra-EU transit trade and differing reporting thresholds (Loschky 2006). In short, mirror discrepancies result from systematic and idiosyncratic drivers, with the latter near impossible to model in a large-n analysis.

Determinants of Asymmetries Analysis

To explore how much candidate drivers of discrepancies that can be well measured or proxied explain asymmetries, we analyze our most fine-grained discrepancy data: dyadic ABBA terms for a substantial cross-section of dyads and more than half a century (1950–2004).[13] We estimate the model

$$Y_{i,t} = \lambda_0 + \lambda_1 \cdot S_{i,t} + \mu_i + \delta_t + \nu_{i,t}$$

where i denotes a (unidirectional) dyadic trade flow and t years. Our dependent variable $Y_{i,t}$ takes the log value of the absolute ABBA discrepancy for a dyad-year in constant 1967 USD. $S_{i,t}$ is a vector of independent variables. μ_i are dyad-fixed effects to absorb the influence of factors constant within dyads over time. Year-fixed effects δ_t control for temporal shocks affecting all dyads simultaneously. $\nu_{i,t}$ is the error term. Standard errors are robust clustered at the dyad level.

We compare the model fit and explained variance for several specifications. $S_{i,t}$ in model 1 only includes the log of the mirror-average trade volume (in constant USD). Model 2 adds average dyad-specific c.i.f. conversion rates computed by the OECD,[14] dummies equal to 1 if both dyad countries are OECD, respectively both non-OECD economies, to evaluate the role of economic development, as well as proxies for similarity of dyads—in geographic, political, and cultural terms; EU membership; and democracy—while avoiding multicollinearity. We also include a dummy for trade flows involving at least one oil export-dependent economy,[15] as well as those involving five well-known entrepot trade jurisdictions,[16] and a dummy for China, whose data are frequently portrayed as particularly unreliable. Model 3 includes year-fixed effects δ_t; model 4 adds dyad-fixed effects μ_i. In separate analyses (online appendix table B1), we examine tariff rates and capital account openness, available only for smaller subsets of our sample.

Table 1 summarizes the main results. Unsurprisingly, a trade flow's size is a powerful predictor of the size of a discrepancy. Notably, c.i.f. conversion rates per se do not appear to drive asymmetries significantly. Dyads of less developed states tend to have larger discrepancies than developed ones. Model 2 shows that countries further removed from one another geographically, culturally, and politically tend to report higher discrepancies. The same is true for dyads involving island states, landlocked states, and countries with large territories. Echoing previous studies, more democratic countries and dyads yield slightly smaller

[10] Interview with senior trade statistician at OECD Statistics Directorate, Paris, June 6, 2017.

[11] Interview with senior trade statistician at OECD Statistics Directorate, Paris, June 6, 2017; interview with senior WTO statistician, Geneva, August 22, 2017; interview with IMF statisticians, Washington, DC, September 19, 2017.

[12] Personal communication with US Bureau of Economic Analysis officials, March 18, 2017.

[13] We rely primarily on the Tomz dataset on trade flows, based on IMF DOTS, covering dyads over a 50-plus year period and including many relevant explanatory and control variables. We add information on mirror trade flows, derived from IMF DOTS, and on additional explanatory variables.

[14] The conversion rates are from 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 1995–2014. We use authors-provided dataset with product-weighted dyad-level annual c.i.f. rates, and calculate 1995–2014 period averages for each dyad, which we treat as the "best guess" for c.i.f. rates for our longer time period.

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

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

Table 1. Sources of ABBA-measured mirror discrepancies

Dependent variable (DV): absolute mirror discrepancy (log, constant USD)	*(1)*	*(2)*	*(3)*	*(4)*
Trade volume	0.84	0.89	0.91	0.94
	(264.6)	(243.77)	(251.998)	(247.04)
C.I.F. rate (dyad mean)		−0.07	−2.20	
		(−0.11)	(−3.31)	
Distance		0.17	0.18	
		(10.62)	(11.02)	
Shared border		0.11	0.06	
		(1.58)	(0.88)	
Number of landlocked in dyad		0.21	0.12	
		(10.16)	(5.67)	
Number of island states in dyad		0.14	0.16	
		(5.17)	(5.72)	
Land area (product)		0.02	0.05	
		(3.24)	(9.05)	
GDP (product)		−0.03	−0.10	−0.17
		(−4.02)	(−13.67)	(−11.17)
Both industrial states		−0.34	−0.12	
		(−7.44)	(−2.56)	
Both nonindustrial states		0.41	0.29	
		(15.85)	(11.03)	
Polity IV score (product)		−0.02	−0.01	0.002
		(−2.15)	(−0.81)	(0.20)
Both formal GATT/WTO members		0.06	−0.06	0.01
		(2.75)	(−3.19)	(0.58)
Reciprocal PTA in force		−0.20	−0.27	−0.19
		(−7.90)	(−10.43)	(−6.52)
Common currency		0.12	0.13	−0.12
		(1.28)	(1.37)	(−0.67)
Both EU members		0.43	0.37	0.05
		(5.13)	(4.49)	(0.66)
Common colonial orbit		−0.54	−0.23	
		(−2.13)	(−0.91)	
Common language		−0.14	−0.17	
		(−4.05)	(−4.80)	
Oil exporter		0.10	0.05	
		(3.19)	(1.51)	
Entrepot trade hub		0.12	0.14	
		(3.26)	(3.75)	
China dummy		0.04	−0.08	
		(0.72)	(−1.33)	
Year-fixed effects?	No	No	Yes	Yes
Dyad-fixed effects?	No	No	No	Yes
Number dyads	10,457	9,852	9,852	9,852
N	195,457	188,499	188,499	188,499
R^2	0.67	0.68	0.68	0.75
AIC	772,140	739,498	736,142	694,117
BIC	772,160	739,711	736,356	694,736

Notes: *t*-statistic is given in parentheses. Dyad-clustered robust standard errors. Constant omitted from output.

mirror discrepancies (Hollyer, Rosendorff, and Vreeland 2011). Preferential trade agreements correspond to smaller discrepancies, whereas General Agreement on Tariffs and Trade (GATT)/WTO membership yields mixed results. Entrepot and oil trade yield higher discrepancies. The China dummy is insignificant.

Counterintuitively, 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 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 adding controls improves model performance, with the full fixed-effects model 4 performing best. The complementary analysis in online appendix table B1 suggests that capital openness is associated with smaller discrepancies, and higher tariff rates with larger discrepancies—in line with expectations of deliberate over- and under-invoicing. However, these relationships are statistically insignificant when including other controls.

All that said, the most striking result is how little variation the various explanatory variables account for—even in the full fixed-effects model (model 4). The size of trade flows does the most explanatory work—neither surprising nor particularly elucidating. Absolute trade volumes alone account for 67 percent of variation. Adding all other variables, or year-fixed effects, barely improves model fit (see

the 0.68 R^2 in models 2 and 3). Also, the inclusion of full dyad-fixed effects (model 4) has little effect on R^2 (0.75 in model 4).[17] Additional analyses, outlined in the online appendix, yield comparable results.[18]

These analyses underline that the discrepancies are highly idiosyncratic, hard to identify and to control for empirically. We do not know how much ABBA terms reflect multiple, layered biases versus unsystematic error, and we cannot assume that errors are randomly distributed and therefore cancel each other out at the aggregate level.

The Handling of the Mirror Problem in Existing IR Scholarship

In economics, some datasets have been developed to (partly) address the mirror problem, including the Global Trade Analysis Project (GTAP) (Gehlhar 1996), BACI (Gaulier and Zignago 2010), and OECD BIMTS (Fortanier and Sarrazin 2016) databases. Notwithstanding minor differences in methodology,[19] 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 IR users in mind: they cover only subsamples of countries and (particularly) short time periods, and they are designed primarily for dyadic-product-level analyses rather than country-dyads.[20] Indeed, only 1 of the 108 papers reviewed above uses one of those datasets. The methods developed in these databases, furthermore, cannot be fully extended to other country-dyads, years, or products, since the methods and coding used to generate inferred reporter reliability are not publicly available. We therefore develop, below, our own approach to such balancing to the extent that current data allow.

How IR studies *do* "address" mirror discrepancies is minimalist and problematic. While IMF DOTS provides both-sides-of-the-mirror data, most studies in IR and political science use the import values, either consciously or by using the major off-the-shelf datasets in IR (e.g., COW or Gleditsch). Values from partner countries' export statistics are disregarded. Researchers sometimes justify this practice arguing that authorities have greater incentives to monitor imports than exports for the collection of customs duties. *Ceteris paribus* import data should be better.

Several factors may argue in favor of export statistics, however, at least sometimes. First, in trading relationships in which exporting countries are the ones with higher statistical capacity than importing trade partners, their records are likely to be more accurate. Second, exporters fulfilling peculiar functions in the international trading system—for example, being an entrepot trade hub, a platform for commodity traders, or an oil exporter—often have greater expertise in adjusting their trade statistics to these particularities. Third, growing e-commerce and disintermediation pose a challenge for trade data collection procedures (Weymouth 2017). 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 (Braml and Felbermayr 2019). All else equal, this will make *exporting-country* records more reliable. These dynamics may be particularly important in custom unions—such as the EU, accounting for roughly 15 percent of global trade and free of internal custom inspections—where member governments rely primarily on data from corporate accounts to estimate trade flows (Eurostat 2016).[21]

A priori, then, we see no reason to assume that for any set of mirror values, one side of the mirror 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 work well in total. […] Discrepancies are large. You can't say it's a rounding error.[22]

Suggested Approaches to Better Account for the Mirror Problem in IR Research

The mirror problem operates at various levels, and for the time being there is no one way to solve it. There are, however, ways to better account for it in our analyses by checking the robustness or sensitivity of trade-related findings to measurement problems (Barbieri and Keshk 2011; Boehmer, Jungblut, and Stoll 2011). We focus on the most promising, easily implementable approaches, and these differ depending on the kind of analyses conducted, particularly small-*n* analyses versus large-*n* analyses, and dyadic versus monadic analyses.

Small-n Analyses

When analysts study bilateral trade flows between only two or a small number of countries (say the US–Mexico or Germany–China trade imbalances), it is important to heed the measures provided by both countries (if available) instead of just relying on one. Where similar patterns emerge using either side of the mirror, they are less likely to be measurement artifacts. If, however, discrepancies are large enough to yield different outcomes, inferences should be adjusted accordingly. The symmetry-weighted average values of bilateral trade that we describe in the subsequent paragraph can offer a more refined estimate, and in some cases statistical agencies provide information that can clarify reasons for discrepancies. More generally, in such analyses, scholars should consider potential causes of discrepancies and reason-through how they matter for the analysis at hand.

Large-n Dyadic Analyses

In large-*n* statistical analyses, in-depth investigation of asymmetries is infeasible. However, information from mirror

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

[18] We estimate similar models using logged, as well as nonlogged, ratios of the ABBA discrepancy relative to the mirror average of dyadic trade volume as dependent variable (online appendix tables B2 and B3), and rerun the baseline model with trade volumes in current USD (online appendix table B4) and excluding all observations with zero trade reported (online appendix table B5). In models without trade volume on the right-hand side (online tables B2 and B3), R^2 values are as low as 0.19 without fixed effects (0.46 with fixed effects).

[19] A useful overview is provided in Fortanier and Sarrazin (2016).

[20] 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 1994–2018; OECD BIMTS is work-in-progress feeding into the Trade-in-Value-Added (TiVA) initiative, encompassing 120 countries between 2007 and 2016.

[21] Our analysis of trade–statistic discrepancies in the EU-27 in online appendix table B6 shows that within-dyad discrepancies increase as European countries joined the Common Market—an effect driven by deterioration in import records (model 4).

[22] Interview with senior trade statistician at OECD Statistics Directorate, Paris, June 6, 2017.

discrepancies can be leveraged to strengthen the robustness of inferences. Whereas other, more cumbersome strategies are potentially available,[23] we recommend relatively straightforward and more easily implementable research practices.

Most of the statistical analyses of bilateral trade in large samples of country pairs that we reviewed rely on the record provided by importers, effectively ignoring the mirror information one can derive from export records. To evaluate the robustness of trade-related findings in large-*n* dyadic studies, we recommend two robustness checks: a mirror-substitution check of both sides of the mirror and running models with a symmetry-weighted average value of bilateral trade.

These checks only work for the subsample of dyads for which mirror trade information is available (in the global sample, a third of dyads accounting for three quarters of global trade).[24] They amount to a proper robustness check only if there are no meaningful differences between trade-related results in the original sample and in the subsample of observations for which two independently recorded mirror trade flows are available. Where there are such differences, mirror checks constitute sensitivity analyses rather than robustness checks: even if there are selection issues, the recommended procedures can still indicate how sensitive an analysis is to data discrepancies, and the likely direction of biases in results due to trade-related measurement errors—a point we illustrate in the Kastner replication below.

This being said, independently of whether or not the replicability condition holds, mirror checks do not reveal the "true" values of trading flows, or their causes or impacts, which remain unknown. But by providing a plausible upper and lower-bound (mirror substitution) and a "best guess" estimate (symmetry-weighted average) in the sample of dyads with mirror records, our suggested checks substantially reduce *uncertainty about the uncertainty* about trade-related findings.

Mirror Substitution (Upper/Lower Bound)

We encourage scholars working with bilateral trade data to re-run their baseline models in the subsample of dyads with mirror records with import- as well as export-records-based estimates—the "mirror substitution" check. Doing so does not assume or reveal that either of the two values is "correct." But the true value is likely somewhere in between. Comparing import-records to export-records-based results is thus useful because they can offer "upper bound" and "lower bound" estimates of the effects or origins of trade. Inconsistent findings call for further investigation. If a finding holds for either specification, confidence grows that a relationship is not merely a measurement artefact.

Symmetry-Weighted Average ("Best Guess")

In addition to upper and lower bounds, we also recommend scholars re-run their analyses in the subsample of mirror records using a series of symmetry-weighted average values of bilateral trade, which we provide in a database accompanying this article. They provide single trade estimates that explicitly heed mirror discrepancies. Instead of strong assumptions about the ultimate sources of mismeasurement, they focus on the asymmetries a reporting country has with all other country-reporters in a given year. A reporter whose trade values are very different from mirror records from all partner countries can be seen as less reliable than a reporter whose statistics are relatively close to other countries' records.

We thus derive an average weighted by the inferred credibility of each reporter. This measure offers a more plausible "best guess" than mechanical reliance on either side of the mirror alone. In constructing weighted averages we follow the key steps to reconciling mirror statistics developed by the OECD/WTO (Liberatore and Wettstein 2021): the TiVA initiative that generated reconciled mirror average flows at the product-level for (only) the most recent years. We adopt the basic methodology with a view to analyses most frequently encountered in IR: global dyadic or monadic samples with long time series.

Here, we make three important assumptions, namely that actual trade values will frequently lie between the two mirror values, that countries with smaller discrepancies with all other countries produce more reliable data than those with larger discrepancies, and that the subsample with mirror values still has a coverage that is relevant to the analysis.

The weighted averages that we construct 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 14). $w_{a\ t}$ is determined by the median of country *A*'s ABBA discrepancies relative to the combined sum of mirror flows, $median(\frac{|trade_{baA\ t\ f.o.b.} - trade_{baB\ t\ f.o.b.}|}{trade_{baA\ t\ f.o.b.} + trade_{baB\ t\ f.o.b.}})$, 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:

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

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

By way of example, this weighting method assigns for US–Mexico in 2010 a weight of 0.71 to US reporting of

[23] Information on mirror discrepancies could be used, for instance, to generate measures of latent variables with measurement error in structural equation modeling, or for Monte Carlo simulations in a given research context, or for the calculation of standard errors in such contexts.

[24] Supplementary analyses (online appendix C) show that the likelihood that two independent mirror records exist is higher for dyads that trade more with each other, involve larger, richer trading partners, and countries with closer economic cooperation (e.g., GATT/WTO membership, reciprocal preferential trade agreement (PTA), or a common currency arrangement). The results also suggest that independent mirror records were marginally more frequent in earlier decades, when country coverage was more limited and IMF data imputation less common. To the extent that these factors correlate with relatively higher statistical capacity, the checks naturally focus on the subsample with above-average measurements. Results that do not hold in the subsample of dyads with mirror records are, from that perspective, unlikely to be reliable in a subsample of dyads without mirror records.

US–Mexico flows on average (and 0.29 to Mexican reporting of US–Mexico flows on average), such that the 2010 weighted value of US imports from Mexico is 229,142 (current millions USD) rather than the US-reported 225,235 or Mexico-reported 238,684.[25] This weighting method assigns a weight for import records in our global sample that averages 0.54, with bottom and top quartiles being assigned weights of, respectively, below 0.4 and above 0.7. In more than half of the weighted averages generated (56.6 percent), import records appear more reliable based on symmetry; in 211,488 cases (42.7 percent) export-record based ones (in the remaining 3,580 dyads, both reporters receive equal weights). Other weighting approaches are conceivable, but this best-guess approach can gauge more reliably size and direction of trade-related coefficients than relying on either side of the mirror records alone.

LARGE-n MONADIC ANALYSES

In monadic settings, the mirror problem is harder to track because the source data do not directly offer mirror values. Still, since monadic data are central to many IR analyses, we propose controlling for mirror error through what we call *monadic ABBA terms*. They can be included as "control" variables and analyzed to visualize the interaction between key explanatory variables and ABBA-proxied measurement uncertainty. *Monadic ABBA terms* measure the difference between (1) the sum of the value of all import [export] flows recorded by the reporting "home" economy and (2) the sum of the value of all mirror flows recorded by *partner* countries in the subsample of dyads with mirror information. The basic *monadic ABBA term* can be defined as follows:

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

$$\left| \sum_{i=1}^{n} trade\ i_{baA\ t} - \sum_{i=1}^{n} trade\ i_{baB\ t} \right|$$

$$i.i.f.\ trade\ i_{ba}\ recorded\ twice\ independently$$

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

$$\left| \sum_{i=1}^{n} trade\ i_{abA\ t} - \sum_{i=1}^{n} trade\ i_{abB\ t} \right|$$

$$i.i.f.\ trade\ i_{ab}\ recorded\ twice\ independently$$

These separate measures are important, because many analyses explicitly focus on either imports or exports. That said, they can be fused in a total-trade monadic ABBA term, which can again be normalized, for example by relating it to total trade or GDP.[26]

Figure A3 in the online appendix illustrates some descriptive data on those monadic ABBA terms. They are remarkably large (the mean is 7.6 percent of a country's GDP and the median 3.1 percent) and not concentrated in any specific region of the world. In our replications, we recommend using these error terms to conduct three complementary robustness checks: adding the monadic ABBA term as a "control" variable; dropping the decile of observations with the highest monadic ABBA terms; and plotting the interactive relationship between the trade variable and the underlying monadic ABBA term. While not "solving" measurement problems, these steps clarify how mirror problems affect findings.

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. Together, however, these checks gauge how mirror problems may influence the statistical relationships of interest.

As supplements to this paper, we make available two public datasets of dyadic and monadic ABBA terms for a large panel of countries between 1948 and 2021, 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.

The Mirror Problem in IR Studies of Trade: Five Replication Analyses

With these approaches, we can probe how the mirror problem plays out for five prominent IR studies about economic interdependence. We explore the sensitivity of their findings to mirror-related data uncertainty and explore avenues to better heed it. We have selected studies that capture diverse uses of trade data, research designs, and topics, employing dyadic as well as monadic setups (table 2). They were not selected, or presented, on the basis of the specific outcomes that the replications have yielded.

For each study, we first replicate the original findings and then compare these to estimation approaches described in section "Suggested approaches to better account for the mirror problem in IR research.".

Replication of Dyadic Studies

Our first dyadic replication concerns a research design investigating how trade with China affects geopolitical alliances. Our second and third reexamine analyses that link GATT/WTO membership to trade flows. In all instances, we follow the same steps: we replicate the original results (model 1); rerun the baseline for the subset of the sample for which two independent mirror records are available to check the impact of observations without mirror records selecting out of ABBA (model 2); replace the import-based records with the corresponding entries in export-based records (the "mirror substitution check"; model 3); and replace trade values with the weighted average of mirror records (model 4).

KASTNER (2016)

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

Kastner's original model is a cross-sectional multinomial logit (Kastner 2016, 992–94). The dependent variable is

[25] Such weighting is significantly less skewed than in 1970 (0.81 for the United States and 0.19 for Mexico). See online appendix table D1 for detail, based on three bilateral trade flow examples.

[26] Note that normalizing by trade or GDP (which includes the value of net exports as estimated by the "home" economy) can introduce trade measurement issues in the denominator, however.

Table 2. Selection of studies for replication

	International security/politics	*Political economy*
Dyadic	Kastner (2016) (effect of trade on security diplomacy)	Rose (2004)/Goldstein, Rivers, and Tomz (2007), (GATT/WTO membership affecting trade)
Monadic	Barbieri and Reuveny (2005) (trade affecting violent conflict)	Garrett and Mitchell (2001) (trade affecting welfare states)

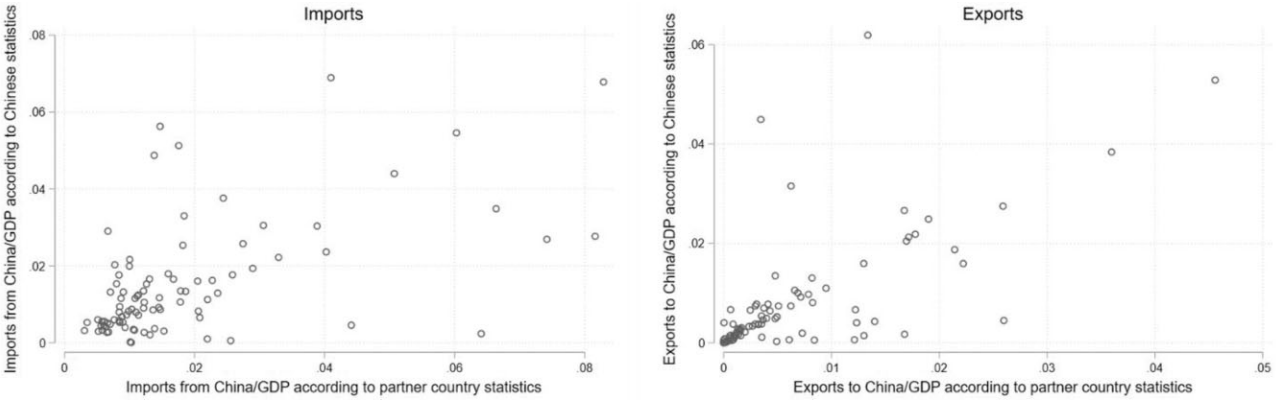

Figure 4. Value of bilateral trade flows with China as a share of GDP in mirror statistics, 2004.
Notes: Observations for which the IMF indicates partner records imputations are excluded. For better readability, both axes in both graphs are truncated at 0.1.

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	0.78	1.96	1.46
	(3.56)	(2.61)	(3.42)	(3.17)
Strong support				
Imports from China/GDP (ln)	0.82	0.15	1.28	0.55
	(3.41)	(0.58)	(2.42)	(1.16)
Control variables as in original?	Yes	Yes	Yes	Yes
N	146	96	96	80
Log-pseudolikelihood	−105.3	−65.5	−61.7	−52.2

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

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 trade dependence, controlling for geographic distance, measures of authoritarianism, security relations with the United States, and national power. All data are from 2004. Trade dependence is operationalized as the foreign governments' bilateral imports from [exports to] China as a share of GDP, as well as their value relative to total imports [exports].[27] For both import and export values, Kastner relies on Chinese data.

Figure 4 illustrates descriptively the mirror problem in this setup. The left-hand scatterplot compares mirror values for imports from China as a share of importer's GDP (Chinese-reported figures on the *y*-axis; partner-country figures on the *x*-axis). The right-hand plot does the same for exports. Both values are from 2004, the year before declaration of the Anti-Secession Law—the case that we reanalyze.[28] The graphs show that mirror discrepancies can be large. The correlations for import mirror records are 0.71 on a linear scale and 0.52 for its log transformation; they stand at 0.96 (linear) and 0.82 (logged) for exports.

Turning, then, to the actual replication, table 3 summarizes the imports-based analyses (full results are available upon request). Model 1 re-establishes the original results. Model 2 restricts 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.

The robustness tests strengthen the original findings. Replicating the original model in the subsample with two independent mirrors reduces the sample substantially (model

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

[28] 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.

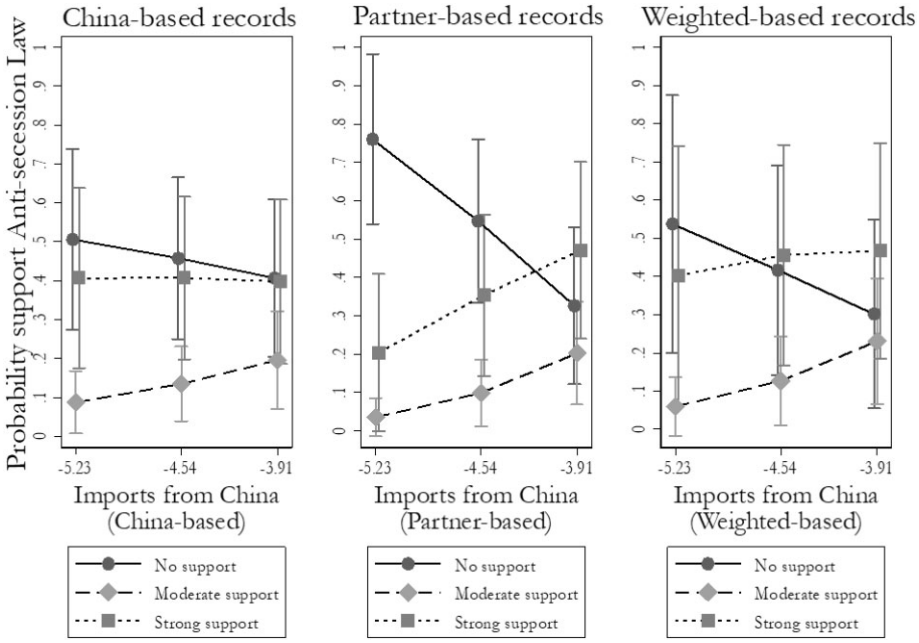

Figure 5. Replication of Kastner (2016).

2 versus model 1), decreasing the size of estimated coefficients in particular for strong support, which loses statistical significance. However, switching from Chinese to partner-country records in this subsample (model 3 versus model 2) leads to similarly substantial jumps in the size and significance of coefficients: 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 versus 2). These differences are substantively modest, as summarized in figure 5: China-based records (model 2 in table 3, first schedule in figure 5) predict 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). Partner-based records (middle schedule in figure 5, model 3 in table 3), in contrast, predict a bigger decrease from 74 to 33 percent. Using weighted averages of import measures (rightmost schedule in figure 5) also strengthens the relationships compared to the Chinese data, although more modestly so. While the replications are more consistent for moderate support (where subsample selection issues are less of an issue), the direction of bias is similar: mirror partner-country records, and weighted-based records, indicate a stronger effect of trade dependence compared to the original study's Chinese-based records. These patterns strengthen our confidence in the original study's positive correlation and indicate that the reported coefficients likely represent lower-bound estimates.

Altogether, Kastner's original findings "pass" the ABBA sensitivity checks. This is important given possible publication bias against "modest" findings. Many past analyses may 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 more readily publishable) findings. The mirror problem thus shapes not only what we do think to know about trade, but also what we think *not* to know (type II error).

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

Our second set of replications concerns 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 negative—effects of GATT/WTO membership on bilateral trade volumes (Rose 2004). Goldstein, Rivers, and Tomz (2007; henceforth GRT) challenged this result. The disagreement centered on two issues: Rose conducted cross-sectional analyses, focusing on between-country variation, while GRT analyzed within-effects over time within a given country. Also, Rose classified country membership by formal participation in GATT/WTO, while GRT considered more fine-grained categorizations accounting for not only *de jure* but also *de facto* ("informal") participation in the regime by some countries (e.g., former colonies).

We use the dataset provided by Tomz and add the mirror information from the IMF DOTS database. We drop the observations that are either missing or outliers for which the log difference in import-based records is greater than 1, leaving 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 yields 183,647 dyad-years with two independently recorded values. The dependent variable used in the analyses is the log value of bilateral trade flows in 1967 US dollars.

We first replicate Rose's between-effects model (summarized in table 4; full results are available upon request) and then GRT's within-analysis (summarized in table 5). Models 1 and 2 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, the implications of the mirror problem are stark, as alluded to in Linsi and Mügge (2019, 370). The import figure subsample with two independent mirror records corroborates Rose's negative relationship between formal GATT/WTO membership and bilateral trade (model 2 in table 4). Although the sample is reduced by about a third, the coefficients are very similar to those in the full sample, fulfilling the replicability condition. However, the coefficients become strongly *positive* and statistically significant once we use the corresponding export figures (model 3 in

Table 4. Replication of Rose (2004)

DV: bilateral trade	*(1)*	*(2)*	*(3)*	*(4)*
Model	*Original baseline (excluding large differences DOTS versus GRT)*	*Baseline in sample with two indepen-dent mirror records*	*Mirror substitution check*	*Weighted mirror average*
Side of mirror	*Import records*	*Import records*	*Export records*	*Average*
Both formal members	−0.10	−0.15	0.56	−0.13
	(−3.16)	(−3.88)	(5.53)	(−3.22)
One formal member	−0.20	−0.16	0.48	−0.07
	(−6.58)	(−4.31)	(4.61)	(−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 is given in parentheses.

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

DV: bilateral trade	*(1)*	*(2)*	*(3)*	*(4)*
Model	*Original baseline*	*Baseline mirror sample*	*Mirror substitution check*	*Weighted mirror average*
Side of mirror	*Import records*	*Import records*	*Export records*	*Average*
Both formal members	0.35	0.26	1.31	0.34
	(8.22)	(4.99)	(7.82)	(6.58)
One formal member	0.18	0.12	1.10	0.21
	(4.73)	(2.47)	(7.01)	(4.25)
Formal and nonmember participant	0.36	0.28	0.96	0.28
	(7.74)	(4.93)	(5.19)	(4.87)
Both nonmember participants	0.45	0.30	−0.10	0.16
	(4.48)	(2.24)	(−0.18)	(1.11)
One nonmember participant	0.08	0.10	0.49	0.12
	(1.53)	(1.47)	(2.25)	(1.78)
Reciprocal PTA in force	0.35	0.36	0.28	0.32
	(14.76)	(12.93)	(3.63)	(12.02)
Nonreciprocal PTA in force	−0.05	−0.07	−0.21	−0.01
	(−1.37)	(−1.80)	(−1.96)	(−0.24)
GSP	−0.16	−0.14	−0.20	−0.13
	(−7.57)	(−6.01)	(−2.87)	(−5.60)
Common currency	0.52	0.42	0.46	0.41
	(5.69)	(5.08)	(1.63)	(4.49)
Common colonial orbit	0.12	0.45	0.64	0.44
	(0.32)	(5.22)	(1.52)	(3.98)
GDP (product)	0.66	0.74	0.80	0.75
	(51.26)	(45.24)	(16.43)	(45.63)
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: Dyad-clustered robust standard errors; *t*-statistic is given in parentheses.

table 4). Formal GATT/WTO members appear to trade *less* than nonmembers if we use import records, but they trade *more* if we use export figures. If we plug in 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 5 percent level.

The results are equally remarkable for the GRT replications, summarized in table 5. They corroborate the GRT claim of a positive GATT/WTO membership effect. The mirror substitution check (model 3 versus 2 in table 5) shows this effect to be several times larger once we use export-based data.[29] Estimating the model with the weighted mirror average, the results are substantively and statistically stronger than in the import-based baseline for

[29] Detailed data analysis shows these large differences to be driven by large discrepancies in dyads in which one country reports zero trade while the other does not. Note that dyads with missing values or mirror-imputed flows are excluded

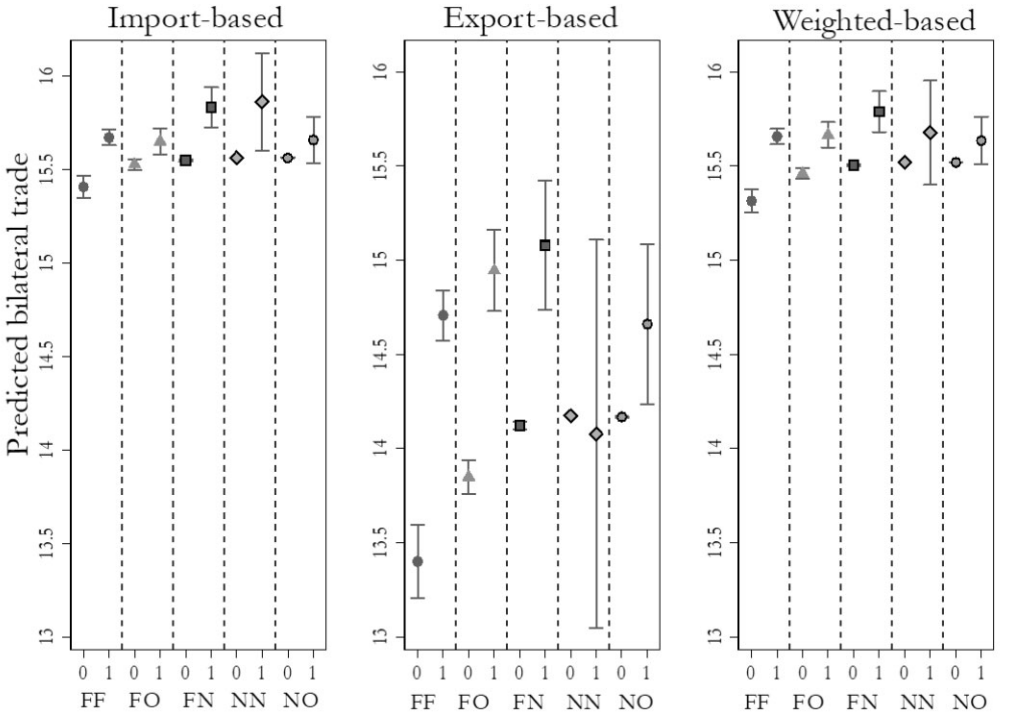

Figure 6. Replication of Goldstein, Rivers, and Tomz (2007).
FF = Both formal members; FO = One formal member; FN = Formal and nonmember participant; NN = Both nonmember participants; NO = One nonmember participant.

formal GATT/WTO membership, but somewhat weaker for "nonmember participants" (countries' accession to "informal" membership)—the theoretical core of GRT's article.

Figure 6 clarifies the substantive meaning of these results, focusing on predicted effects of different GATT membership constellations on bilateral trade levels (on the log scale ranging in the full sample from 13 to 24.5). It shows predicted trade levels based on values of the five GATT/WTO participation constellations in models 2 (import-based), 3 (export-based), and 4 (weighted-based). Holding all other parameters at their means, the implications of these robustness checks differ by GATT membership constellation. Export records indicate stronger trade-enhancing effects for formal GATT membership than import records, but no effect for dyads with two nonmember participants. The weighted averages tend to align more with import-records-based results, with somewhat stronger effects for full members but weaker ones for nonmember participants.

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 other, originally reported substantive and statistical disagreements. Also important, however, attention to the mirror discrepancies in trade data reveals that "informal" nonmember participation plays a smaller role for the discrepant findings than previously estimated.[30]

from the sample, so that these values refer to actually reported zeroes. Also, note that transformation of trade flows in dollar units to logarithmic scale compounds these issues in the setup.

Extensions to Monadic Studies

The mirror problem is essentially a dyadic phenomenon that can be directly explored as above. However, the mirror problem may also matter at the monadic level and can be explored indirectly (more imperfectly) using mirror statistics. The two replications we present illustrate easily implementable approaches to do so. They again cover different IR topics and research designs: the first study assesses how trade openness affects the risk of civil wars in developing countries and the second analyses the link between trade and government spending in advanced industrial economies.

Our main replications pursue the following procedure: we re-establish the original results (model 1); rerun the baseline for the sample for which monadic ABBA terms are available (model 2); include the monadic ABBA term as a "control" (model 3); and rerun the baseline in a restricted sample that excludes the decile of observations with the largest mirror discrepancies (model 4). Finally, we interact the monadic ABBA term with the explanatory trade variable to visualize how measurement errors affect statistical findings.

Barbieri and Reuveny (2005)

Monadic trade data are central to studies linking economic openness to the risk of civil wars. Barbieri and Reuveny's (2005) systematic investigation assesses various globalization measures (trade openness, foreign direct in-

[30] Since the import-based models confirm the original positive correlation for nonmembers, which only weakens markedly when switching the mirror—while strengthening them for formal members—the difference does not appear to be driven by subsample selection alone.

Table 6. Replication of Barbieri and Reuveny (2005)

DV: civil war presence	*(1)*	*(2)*	*(3)*	*(4)*
Model	*Original baseline*	*Baseline merchan-dise 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	−0.009	−0.010
		(−2.33)	(−0.90)	(−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
Pseudolikelihood	−232.9	−183.0	−182.3	−169.0

Note: Robust standard errors clustered by country; *z*-statistics is given in parenthesis.

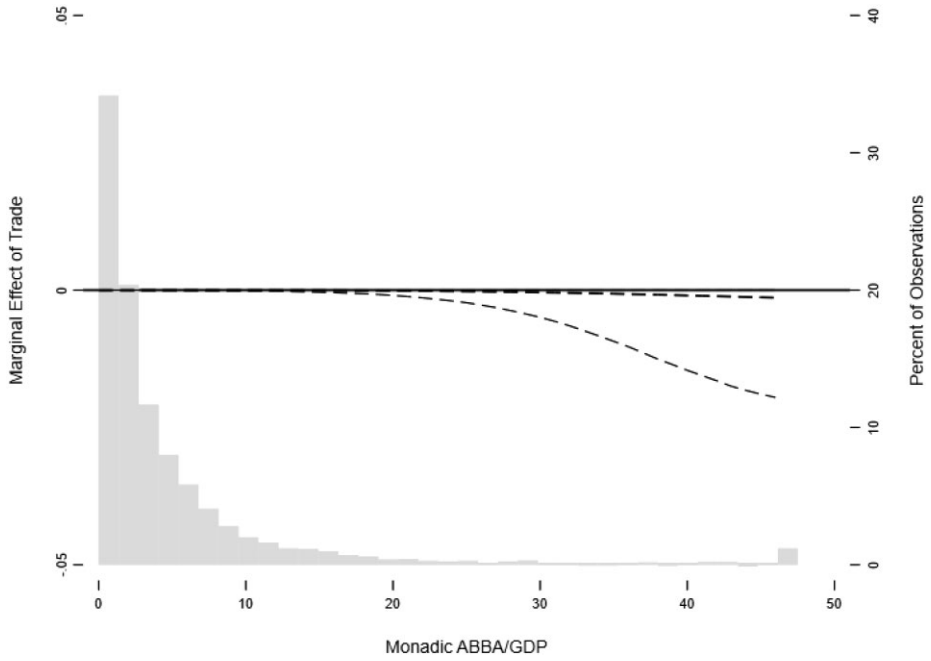

Figure 7. Effect of a one unit increase in trade/GDP on the probability of civil war presence at different ABBA levels.
Notes: For better readability, the maximum for the ABBA term was fixed at its 99th percentile in underlying regressions. All other variables are set at median value. Dotted lines indicate 95 percent confidence interval.
Source: Graph code from Berry, Golder, and Milton (2012).

vestment (FDI) inflows, portfolio capital 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 their duration. Thanks to data provided by the authors, we could reproduce the original results exactly. Our replications estimate the effects of trade openness once we appreciate data problems.

Table 6 summarizes the behavior of the trade variable (full results are available upon request). 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.

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 fails 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).

Figure 7 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 war presence, therefore, may be driven by a modest number of

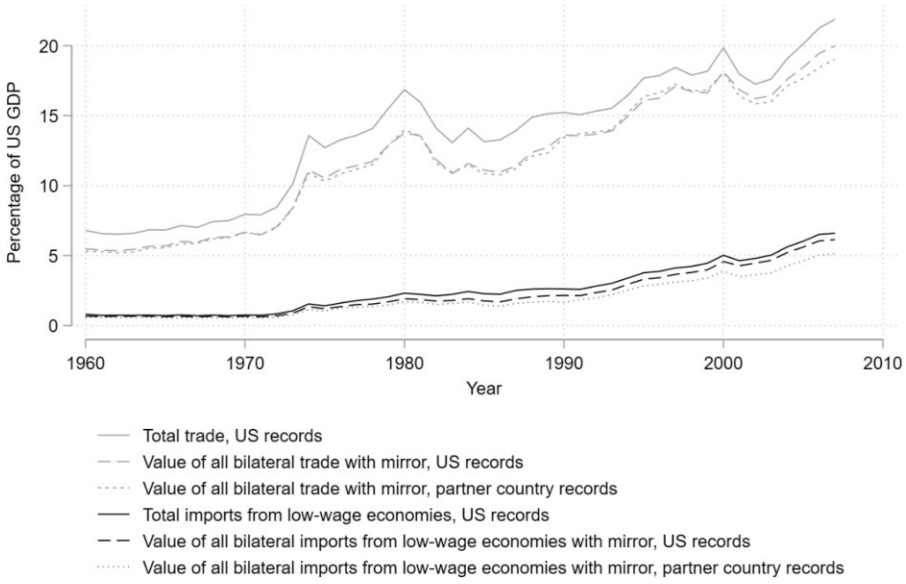

Figure 8. Illustration of ABBA factor and low-wage ABBA factor for the United States.

observations for which mirror discrepancies are very high. And forces triggering violence may spawn inaccurate statistics. In this sense, our results need not invalidate the original findings or the theoretical argument informing the work. Nonetheless, the replication highlights how questionable trade statistics may complicate statistical study of the trade and conflict relationship (Schultz 2015).

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 Garrett and Mitchell's (2001) widely cited study. They investigate how globalization affects welfare states. We concentrate on their analysis linking trade to total 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* (see also Burgoon 2001).

With a dataset provided by Busemeyer, we follow Garrett and Mitchell's research design as closely as possible. We undertake a few modifications to illustrate the effect of trade–data quality: we focus only on trade (not FDI and portfolio flows) in the post-1980 period of interest in the original studies. 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.

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-Organization of the Petroleum Exporting Countries economies. Figure 8 illustrates the resulting two monadic ABBA terms for the United States graphically, with the discrepancies displayed between the two lower lines. Both grow over the decades, and US figures typically outstrip those of its partner countries.

Table 7 summarizes our replications. We re-establish the original baseline (model 1) and the baseline with merchandise (rather than total) trade (model 2); we include the monadic ABBA term as a control (model 3) and rerun the baseline in a restricted model (model 4) that excludes the decile of country-years with highest ABBA terms (in the OECD sample, these are countries with monadic ABBA terms exceeding 6.1 percent of GDP). Finally, we interact the trade variable with the underlying ABBA terms.

Remarkably, the ABBA robustness checks pull in different directions for the two trade measures: the negative relationship between total trade and public spending waxes and the positive effect of low-wage imports wanes. Such waxing and waning, however, does not entail meaningful changes in the levels of statistical significance of the trade parameters.

An interaction between the relevant ABBA terms and the measures of trade (left-hand panel of figure 9) or low-wage trade (right-hand panel) clarifies this pattern. For the total trade variable (left-hand panel), the negative relationship is significantly negative when measurement error due to mirror discrepancies is small. For low-wage imports (right-hand panel), in contrast, the positive relationship is strongest for countries with large ABBA factors; it is small in substantive terms for higher data-quality observations. In the latter case, then, low-quality data seem to upwardly bias the original estimates of the relationship.

Taken together, measurement errors likely alter the modeled relationships between trade and welfare spending. Our replications suggest that the original studies may have underestimated 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 go against the attenuation biases suggested by earlier replication studies focused on estimators and error correction (e.g., Kittel and Winner 2005).

Implications and Conclusion

IR scholarship has hitherto ignored or downplayed mirror discrepancies in trade data. Our analyses yield three analytical insights and two recommendations. First, we have

Table 7. Replication of Garrett and Mitchell (2001)

DV: total spending/GDP *Model*	*(1)* *Baseline*	*(2)* *Baseline merchandise trade*	*(3)* *Monadic ABBA as control*	*(4)* *Censoring ABBA top decile*
Total trade/GDP ($t-1$)	−0.10			
	(−2.69)			
Total merchandise trade/GDP ($t-1$)		−0.15	−0.15	−0.20
		(−2.18)	(−2.17)	(−2.82)
ABBA factor ($t-1$)			0.01	
			(0.10)	
Low-wage imports/GDP ($t-1$)		0.42	0.45	0.39
		(1.83)	(1.37)	(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	1981–1994	1981–1994	1981–1994	1981–1994
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 is given in parentheses.

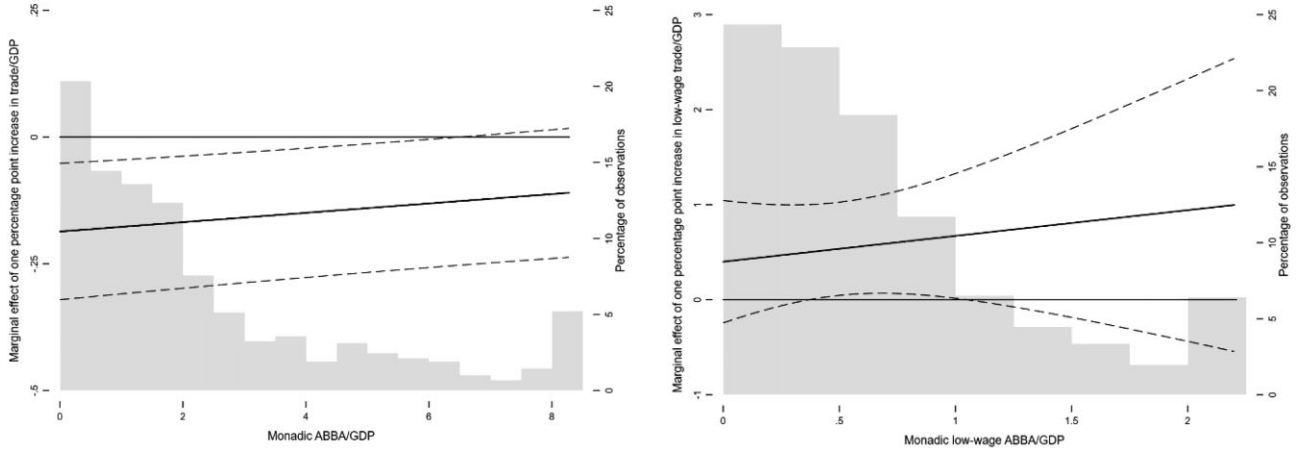

Figure 9. Marginal effect of total trade (left) and imports from low-wage economies (right) on social spending at different values of data quality.
Notes: For better readability, the maximum values for the ABBA terms are fixed at their 95th percentile in underlying regressions. Dotted lines indicate 95 percent confidence interval.
Source: Graph code from Berry, Golder, and Milton (2012).

quantified the gaps between any two countries' estimates about their bilateral trade to construct ABBA terms as proxies for error in the data. Both specific cases, such as US–Mexican trade, and large-*n* analyses of such ABBA terms reveal substantial uncertainty in trade data. Unreflective choice for either import or export data is thus problematic: neither is consistently and obviously superior to the other. It is preferrable to use the information contained in both figures judiciously, following strategies reiterated below.

Second, we have investigated the origins of mirror discrepancies. If we could systematically account for discrepancies, we might control for them. If they were completely random, we could dismiss them as mere data noise. Neither approach, alas, fits our findings. Case studies 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. At the same time, national and dyadic idiosyncrasies can play a big role. Because these factors simultaneously confound trade data, we cannot fully disentangle their resulting discrepancies. Biases in trade data therefore resist eradication. Statisticians try, for example, through bilateral reconciliation exercises or the OECD's TiVA database. However, 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 (Mügge and Linsi 2020).

Third, mirror discrepancies color 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; in some cases, it can reverse their direction. In all studies that we have re-examined, taking mirror discrepancies into account has added nuance to previous empirical findings.

This brings us to several recommendations. First, IR 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 and how they affect the phenomenon under investigation.

Second, matters are less straightforward for larger-*n* comparisons, but our replications suggest several easily implementable approaches to gauge the robustness of inference (cf. Neumayer and Plümper 2017). They include decomposition and remeasuring trade relationships through "mirror substitution checks" and 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 straightforward, 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 both the dyadic and monadic ABBA measures derived from IMF DOTS for a large swath of countries from 1948 to 2021, which will be periodically updated as new data become available. Even though we have limited our examples to IR 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 use trade data.

Beyond the specific issues with trade data, IR scholarship should take measurement problems in political economy more seriously in general. We have here portrayed mirror discrepancies mostly as a "problem." However, they can also be understood as an opportunity, because they can help us reduce uncertainty about the uncertainty of trade measurement. Our study is thus an encouragement to seek out such alternative sources of data covering similar phenomena and use statistical forms of triangulation between them as tools to strengthen our inference, or understand better where data defects preclude strong conclusions.

To be sure, for almost all other statistical data used in IR and beyond, we do not have the luxury of having the same transaction being recorded twice. Furthermore, of the different quantities tracked in BOP data—including services trade, foreign direct investment, and portfolio investment—merchandise trade is arguably the most reliable (Lipsey 2006; Damgaard and Elkjaer 2014; Kerner 2014; Linsi and Mügge 2019). 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, and many aspects of international political life more broadly. It is high time for critical discussion, also in academic training, of data quality and measurement problems in official statistics. At stake is the basic quality of what we know and argue about international economic relations.

Supplementary Information

Supplementary information is available at the *International Studies Quarterly* data archive and the "Mirror Trade" dataverse.

References

Baccini, Leonardo, Andreas Dür, and Manfred Elsig. 2018. "Intra-Industry Trade, Global Value Chains, and Preferential Tariff Liberalization." *International Studies Quarterly* 62 (2): 329–40.

Barbieri, Katherine, and Rafael Reuveny. 2005. "Economic Globalization and Civil War." *Journal of Politics* 67 (4): 1228–47.

Barbieri, Katherine, and Omar Keshk. 2011. "Too Many Assumptions, Not Enough Data." *Conflict Management and Peace Science* 28 (2): 168–72.

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): 471–91.

Berry, William D., Matt Golder, and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *The Journal of Politics* 74 (3): 653–71.

Bhagwati, Jagdish. 1964. "On the Underinvoicing of Imports." *Bulletin of the Oxford University Institute of Economics & Statistics* 27 (4): 389–97.

———. 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): 61–77.

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): 145–67.

Braml, Martin T., and Gabriel J. Felbermayr. 2019. "The EU Self-Surplus Puzzle: An Indication of VAT Fraud?" CESifo Working Paper No 7982, Munich. Last accessed March 16, 2023. https://www.ifo.de/en/node/51244.

Burgoon, Brian. 2001. "Globalization and Welfare Compensation: Disentangling the Ties that Bind." *International Organization* 55 (3): 509–51.

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): 245–60.

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.

Fortanier, Fabienne, and Katia Sarrazin. 2016. "Balanced International Merchandise Trade Data: Version 1." STD/CSSP/WPTGS(2016)18, Paris. Last accessed March 16, 2023. 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). https://one.oecd.org/document/STD/CSSP/WPTGS(2018)11/En/pdf.

Garrett, Geoffrey, and Deborah Mitchell. 2001. "Globalization, Government Spending and Taxation in the OECD." *European Journal of Political Research* 39 (2): 145–77.

Gartzke, Erik, and Quan Li. 2003. "Measure for Measure: Concept Operationalization and the Trade Interdependence-Conflict Debate." *Journal of Peace Research* 40 (5): 553–71.

Gaulier, Guillaume, and Soledad Zignago. 2010. "BACI: International Trade Database at the Product-Level: The 1994–2007 Version." CEPII Working Paper No 2010-23. Last accessed March 16, 2023. 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. Last accessed March 16, 2023. https://www.gtap.agecon.purdue.edu/resources/download/38.pdf.

Gleditsch, Kristian Skrede. 2010. "On Ignoring Missing Data and the Robustness of Trade and Conflict Results: A Reply to Barbieri, Keshk and Pollins." *Conflict Management and Peace Science2* 27 (2): 153–57.

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

Gray, Julia, and Philip B.K. Potter. 2012. "Trade and Volatility at the Core and Periphery of the Global Economy." *International Studies Quarterly* 56 (4): 793–800.

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

International Monetary Fund. 1987. "Report on the World Current Account Discrepancy." Washington, DC.

———. 1993. "A Guide to Direction of Trade Statistics." Washington, DC.

Javorsek, Marko. 2016. "Asymmetries in International Merchandise Trade Statistics: A Case Study of Selected Countries in Asia-Pacific." Statistics Division Working Paper Series SD/WP/02/April 2016. Last accessed March 16, 2023. 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, NY: 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.

Kerner, Andrew. 2014. "What We Talk about When We Talk about Foreign Direct Investment." *International Studies Quarterly* 58 (4): 804–15.

Kim, In Song, Steven Liao, and Kosuke Imai. 2020. "Measuring Trade Profile with Granular Product-Level Data." *American Journal of Political Science* 64 (1): 102–17.

Kim, In Song, Helen V. Milner, Thomas Bernauer, Iain Osgood, Gabriele Spilker, and Dustin Tingley. 2019. "Firms and Global Value Chains: Identifying Firms' Multidimensional Trade Preferences." *International Studies Quarterly* 63 (1): 153–67.

Kittel, Bernhard, and Hannes Winner. 2005. "How Reliable is Pooled Analysis in Political Economy? The Globalization-welfare State Nexus Revisited." *European Journal of Political Research* 44 (2): 269–93.

Liberatore, Antonella, and Steen Wettstein. 2021. "The OECD-WTO Balanced Trade in Services Databae (BPM6 Edition)." Last accessed March 16, 2023. https://www.wto.org/english/res_e/statis_e/daily_update_e/OECD-WTO_Batis_methodology_BPM6.pdf.

Linsi, Lukas, and Daniel Mügge. 2019. "Globalization and the Growing Defects of International Economic Statistics." *Review of International Political Economy* 26 (3): 361–83.

Lipsey, Robert E. 2006. "Measuring International Trade in Services." National Bureau of Economic Research Working Paper Series No. 12271 (May). Last accessed March 16, 2023. https://users.nber.org/~confer/2006/criws06/lipsey.pdf.

Loschky, Alexander. 2006. "Asymmetrien in Der Aussenhandelsstatistik." Last accessed March 16, 2023. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2006/03/asymetrien-032006.html.

Miao, Guannan, and Fabienne Fortanier. 2017. "Estimating Transport and Insurance Costs of International Trade." OECD Statistics Working Papers No. 4.

Morgenstern, Oskar. 1963. *On the Accuracy of Economic Observations*, 2nd ed. Princeton, NJ: 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* 27 (2): 403–27.

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): 452–67.

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

Office for National Statistics. 2017. "Asymmetries in Trade Data: A UK Perspective." London. Last accessed March 16, 2023. https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/articles/asymmetriesintradedataaukperspective/2017-07-13.

———. 2020. "Asymmetries in Trade Data: Updating Analysis of UK Bilateral Trade Data." London. Last accessed March 16, 2023. https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/articles/asymmetriesintradedatadivingdeeperintoukbilateraltradedata/updatinganalysisofukbilateraltradedata.

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: 125–45.

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

Weymouth, Stephen. 2017. "Service Firms in the Politics of US Trade Policy." *International Studies Quarterly* 61 (4): 935–47.

Yeats, Alexander J. 1978. "On the Accuracy of Partner Country Trade Statistics." *Oxford Bulletin of Economics and Statistics* 40 (4): 341–61.

———. 1990. "On the Accuracy of Economic Observations: Do Sub-Saharan Trade Statistics Mean Anything?" *The World Bank Economic Review* 4 (2): 135–56.