INSTITUTE for REPLICATION

No. 23
I4R DISCUSSION PAPER SERIES

New Data, New Results? How Data Sources and Vintages Affect the Replicability of Research

Iasmin Goes

March 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 23

New Data, New Results? How Data Sources and Vintages Affect the Replicability of Research

Iasmin Goes[1]

[1]*Colorado State University, Fort Collins/USA*

MARCH 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

New Data, New Results? How Data Sources and Vintages Affect the Replicability of Research

Iasmin Goes*

March 2023

Abstract

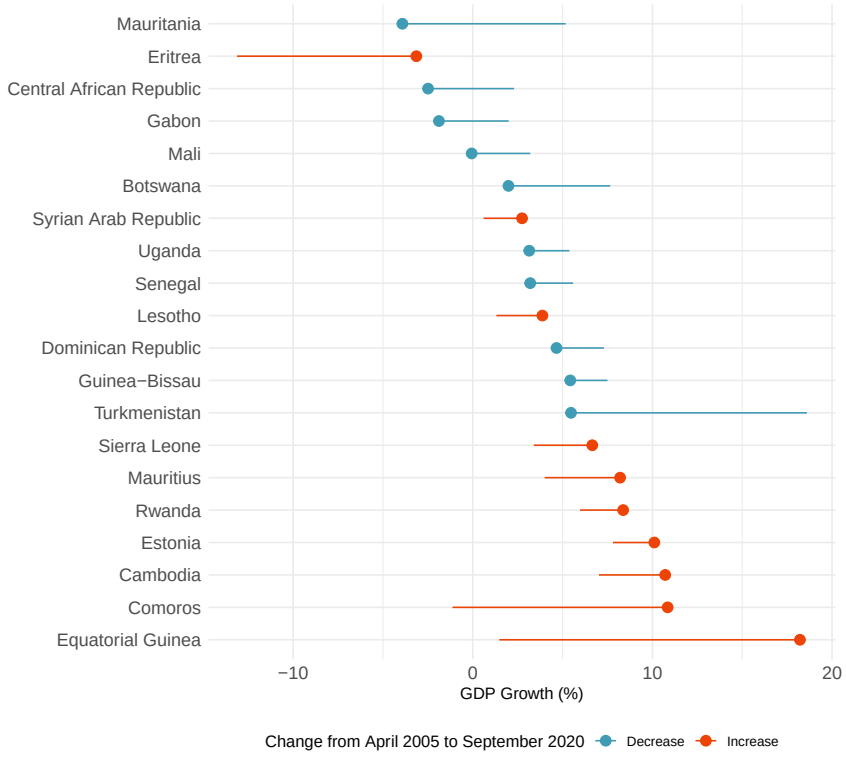
Macroeconomic variables like unemployment, inflation, trade, or GDP are not set in stone: they are preliminary estimates that are constantly revised by statistical agencies. These data revisions, or data *vintages*, often provide conflicting information about the size of a country's economy or its level of development, reducing our confidence in established findings. Would researchers come to different conclusions if they used different vintages? To answer this question, I survey all articles published in a top political science journal between 2005 and 2020. I replicate three prominent articles and find that the use of different vintages can lead to different statistical results, calling into question the robustness of otherwise rigorous empirical research. These findings have two practical implications. First, researchers should always be transparent about their data sources and vintages. Second, researchers should be more modest about the precision and accuracy of their point estimates, since these estimates can mask large measurement errors.

*Assistant Professor, Colorado State University. Contact: iasmin.goes@colostate.edu

1 Introduction

How much did the economy of Equatorial Guinea grow in 2000? Between 2005 and 2020, the World Bank's World Development Indicators (WDI) reported four different values for Equatorial Guinea's 2000 GDP growth, from 1.47 to 18.2%. As Figure 1 shows and Johnson et al. (2013) confirm, this is no exception: researchers seeking to explain economic growth might come to different conclusions depending on their data source and version. As with growth, recent estimates of unemployment, inflation, or trade are preliminary and under constant revision, in what Croushore and Stark (2003) call "data vintaging."

Figure 1: GDP Growth in 2000 for Selected Countries: Difference in Values Reported by Different WDI Releases



This figure examines GDP growth in 2000, comparing values reported by the April 2005 WDI release (the beginning of each line) to values reported by the September 2020 WDI release (the dots). Red lines indicate that the value reported in 2020 increased relative to the value reported in 2005, whereas blue lines indicate that this value declined. The 20 countries included in this figure reported the highest change from 2005 to 2020.

This study examines how data vintaging affects the replicability of published research.

Would researchers come to different conclusions if they used different data? To answer this question, I survey all articles published in a prominent journal between 2005 and 2020, collecting detailed information about data usage, and replicate three articles using various data sources and vintages of the same source. I find that using different data can lead to different statistical results, calling into question the robustness of rigorous research. Researchers should not only disclose their data sources and vintages but also be more modest about the precision and accuracy of their point estimates, which can mask large measurement errors.

2 The Logic Behind Data Vintaging

Datasets like the WDI are revised to incorporate improved data, eliminate errors, and account for new price benchmarks (Ciccone and Jarociński, 2010). The WDI relies on data reported by national statistical agencies; as agencies improve their statistical capacity, they collect and disseminate better data. In 2014, for example, Nigeria's GDP increased by 89% after the government updated the base year for calculations, incorporating information about the telecommunication and film-making industries (Kerner et al., 2017). International organizations advise governments to update their GDP base year at least every 10 years, but half of the 189 countries surveyed by Berry et al. (2018) use older base years, reporting figures that are likely biased downward.

New administrations are often eager to rectify the errors committed by previous administrations. Data manipulation increased under the Kirchner and Rousseff presidencies in Argentina and Brazil, respectively; in both cases, the successor came from a rival party and publicized these data issues (Aragão and Linsi, 2022). Shortly after coming to power, Prime Minister Papandreou requested help from Eurostat and the IMF to revise Greece's public finance statistics, which had long been misrepresented to follow EU rules (Alt et al., 2014). Some countries deliberately report biased figures to seem poorer and meet World Bank eligibility criteria for foreign aid, though these figures are typically corrected *ex post* (Kerner

et al., 2017).

Lastly, the WDI standardizes country-specific data using a purchasing power parity (PPP) adjustment set by the International Comparison Program (ICP) based on international price surveys. Until 1996, these surveys covered developed countries and made extrapolations for the rest of the world; updates in 2005, 2011, and 2017 began to cover developing countries (Deaton and Aten, 2017). Using satellite-recorded nighttime lights as a "true" measure of economic activity, Pinkovskiy and Sala-i Martin (2020) find that vintages based on newer ICP benchmarks have smaller measurement errors.

Data vintaging can be good news. The standardization methodology is improving: current estimates of the past are closer to the "truth" than past estimates of the past. It is better to revise inaccurate figures than to keep them consistently inaccurate. However, data vintaging calls into question the robustness of existing findings, since many vintages provide conflicting information. With few exceptions (Boehmer et al., 2011; Hollyer et al., 2014; Inklaar and Prasada Rao, 2017; Fariss et al., 2022), researchers did not examine these issues until recently.

3 A Survey of Published Studies

To gauge the prevalence of vintaged data in political science, I assembled all 459 research articles and research notes published in *International Organization* between 2005 and 2020, including special issues. Of these studies, 173 conducted a cross-country statistical analysis using contemporary macroeconomic variables, like GDP growth, unemployment, inflation, or trade.[1]

Table 1 classifies each study according to its data sources mentioned in the main text or appendix. The numbers amount to more than 100% because if a single study mentioned multiple sources, I recorded all. This is a conservative estimate, as 24 studies (13.87%) do

[1]The remaining are qualitative studies, surveys, experiments, network analyses, formal models, or statistical analyses that use historical, subnational, firm-level, or conflict data.

not mention *any* source of macroeconomic data. This does not mean that they did not use such data, only that they did not volunteer such information.

Table 1: Relevant Studies Published in *International Organization* According to Data Source, 2005–2020

	Number of studies	Percentage
Use Relevant Data	173	100.00
Mention Data Source	149	86.13
World Development Indicators	106	61.27
Penn World Table	52	30.06
Maddison	9	5.20
Other	25	14.45
Mention Data Vintage	119	68.79
World Development Indicators until 2005	28	16.18
World Development Indicators 2006–2010	23	13.29
World Development Indicators 2011–2015	20	11.56
World Development Indicators 2016–2020	8	4.62
Penn World Table 5.6, 1994	24	13.87
Penn World Table 6.1–6.3, 2002–2009	14	8.09
Penn World Table 7.0–7.1, 2011–2012	7	4.05
Penn World Table 8.0, 2013	2	1.16
Maddison 2003	3	1.73
Maddison 2007	1	0.58
Maddison 2010	4	2.31

Among the studies that mention their data source, 106 use the WDI. Some combine multiple sources. To construct their GDP variable, Goldstein et al. (2007, 50) "turned first to the 2005 edition of the World Bank's *World Development Indicators* ... then extended the series backwards to 1946, using U.S. dollar figures from the Penn World Tables, the United Nations, the Oxford Latin American Economic History Database, and the IMF *International Financial Statistics*. In a few cases, we used the GDP indices from Maddison ... to complete the data set." Mansfield and Reinhardt (2008, 636, footnote 61) use GDP data "from a hierarchy of sources, starting with the World Bank's *World Development Indicators*, the OECD's *Monthly Statistics of International Trade*, UNCTAD's *Handbook of Statistics On-Line*, the IMF's *World Economic Outlook*, the Penn World Table version 6.1, and the IMF's

International Financial Statistics." Yet, WDI, Penn World Table (PWT), and Maddison figures are quite different from each other — even if the underlying data are the same — due to differences in currency conversions or PPP adjustments. Ram and Ural (2014) identify 33 countries for which GDP estimates from the WDI and the PWT differ by at least 25%. It is unclear if authors combining multiple sources are aware of these differences or address potential discrepancies.

Instead of providing a direct source, 21 articles refer readers to Gleditsch (2002), who collected GDP, trade, and population data from PWT version 5.6, imputing missing observations and providing additional estimates from the CIA's World Factbook. Others refer readers to Fearon and Laitin (2003), who used WDI data to extend PWT estimates for GDP growth. However, the data compiled by Gleditsch and Fearon and Laitin rely on older PWT and WDI vintages, which have larger measurement errors. Many authors do not know the true nature of the data underlying their empirical analyses, since they did not collect these data themselves.

Though 86.13% of all articles mention their data *source*, only 68.79% mention their data *vintage*, indicating a release date or version number ("World Development Indicators *2006*" or "Penn World Table *5.6*"). Given the differences between data vintages, this information should always be provided.

4 Empirical Consequences of Data Vintaging

To show that even rigorous research is vulnerable to data vintaging, I use different sources and vintages to replicate three of the surveyed studies, selected following three criteria. First, they had to be transparent about their sources and vintages, allowing me to locate equivalent variables elsewhere. Second, these had to be older studies, ensuring that there were more recent comparable data. Third, supplementary materials (both data and replication code) had to be publicly available, allowing me to estimate the models exactly as published. Since

the *International Organization* website does not provide supplementary materials for issues before 2011, my selection was restricted to authors who provided this information on their own websites.

4.1 De Soysa and Neumayer (2005)

Using data for 135 countries between 1980 and 2000,[2] de Soysa and Neumayer (2005, 732) show that economic globalization has a positive and statistically significant effect on sustainable development, defined as a state's "ability to maintain (increase) the aggregate value of manufactured, human, and natural capital." In eliminating price distortions and promoting an efficient allocation of resources across borders, globalization minimizes waste.

The outcome, genuine savings (% of GNI), combines six WDI variables.[3] The key independent variable, trade (% of GDP), comes from the WDI, as do six control variables: current GNI per capita in PPP; agriculture (% of GDP); GDP per capita growth (%); population; population density; and urban population (% of the total population). The random effects generalized least squares model also controls for regime type, fuel exporter status, political constraints, stability of the political system, occurrence of a currency crisis or a civil war, and number of peace years since 1946.

All WDI variables come from the 2002 release, though data for Angola and Sudan come from the 2003 release "because their values seem to be reported with errors" (de Soysa and Neumayer, 2005, 740, footnote 56). Upon closer inspection of the 2002 data, the genuine savings rate for Angola and Sudan is exceptionally low, but the World Bank (2023) did not flag these countries as problematic, and more recent releases report similarly extreme values. Even if the 2002 vintage suffers from measurement error and the 2003 vintage does not, there is no evidence that this error is limited to Angola and Sudan.

Table 2 compares the original results to results using 2002, 2012, and 2022 WDI data.

[2]Despite its title, the study includes observations for the year 2000.

[3]The genuine savings rate is "equal to net national savings plus current education expenditures and minus energy depletion, mineral depletion, net forest depletion, and carbon dioxide damage" (de Soysa and Neumayer, 2005, 762). Using this formula, I obtain the same values as the authors.

Table 2: The Effect of Trade Dependence on Genuine Savings (Random Effects GLS), 1980–1999

	(1) Original Model	(2) WDI 2002	(3) WDI 2012	(4) WDI 2022
Trade/GDP (ln)	2.416***	2.109**	2.901***	−0.898
	(0.774)	(0.999)	(1.053)	(2.747)
GNI pc (ln)	21.933***	21.070***	12.256**	37.320
	(6.248)	(7.938)	(6.323)	(33.424)
(GNI pc)^2 (ln)	−0.977***	−0.986**	−0.532	−2.617
	(0.370)	(0.473)	(0.371)	(1.895)
Economic Growth	−0.009	0.029	0.115***	0.100
	(0.024)	(0.031)	(0.031)	(0.079)
Agriculture/GDP	−0.052	−0.131**	−0.166**	0.056
	(0.047)	(0.060)	(0.063)	(0.174)
Currency Crisis	0.017	−0.255	1.117**	0.828
	(0.397)	(0.520)	(0.560)	(1.100)
Fuel Exporter	−18.230***	−22.429***	−21.671***	−1.718
	(2.397)	(2.711)	(2.590)	(20.317)
Democracy	1.034	1.009	0.577	1.196
	(0.714)	(0.939)	(0.934)	(2.160)
Political Constraints	−1.203	1.239	−2.597	−4.619
	(1.503)	(1.978)	(1.954)	(3.711)
Government Stability	−0.344	−0.337	−0.175	0.168
	(0.345)	(0.455)	(0.467)	(0.909)
Population Density (ln)	1.078**	1.406***	1.232**	1.741
	(0.435)	(0.499)	(0.490)	(3.627)
Population Size (ln)	0.116	−0.098	0.534	−1.021
	(0.405)	(0.471)	(0.479)	(3.214)
Population Urban	−8.490***	−7.855***	−6.664***	16.062
	(1.524)	(1.816)	(1.783)	(11.964)
Civil War	−1.473**	1.320	−0.038	1.149
	(0.734)	(0.956)	(0.953)	(1.794)
Peace Years	0.010	0.018	0.008	0.032
	(0.025)	(0.032)	(0.031)	(0.091)
Constant	−89.726***	−81.206**	−52.733**	−173.620
	(26.202)	(33.409)	(26.636)	(145.987)
Number of Observations	2,069	2,058	1,822	840
Countries	135	135	122	109

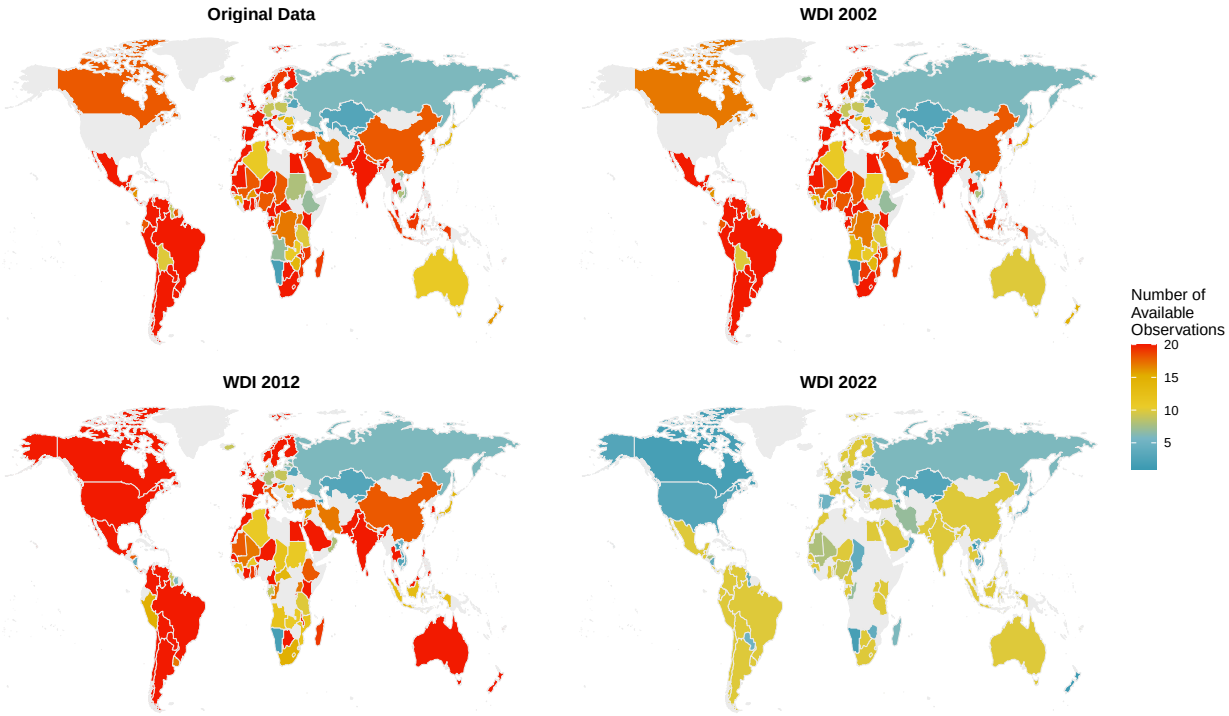
This is a replication of Model 1, Table 1 in de Soysa and Neumayer (2005). Standard errors appear in parentheses. All regressions assume an AR1 correlation structure and include year dummies. All independent variables are lagged one year.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Since older benchmarks tend to be biased downward, I expect to recover larger effect sizes when using newer vintages. But these differences should be in levels, not trends. The effect of

the independent variable on the outcome should be substantively and statistically consistent; both variables come from the same source and should suffer from similar measurement errors.

Models 1 to 3 confirm these expectations and support de Soysa and Neumayer's finding: countries that trade more tend to have a higher genuine savings rate. The coefficient for *Trade/GDP* is smallest when using 2002 data and largest when using 2012 data. Since the exact point estimate is not robust to using different vintages, the interpretation of results should focus on the direction and statistical significance of the effects.

Figure 2: Available Observations by Country, 1980–1999



These maps show the number of complete observations available by country for the variables included in de Soysa and Neumayer's analysis. 30 countries (including Angola and Sudan) are included in the original data, but not in the 2022 WDI; conversely, the 2022 WDI includes Israel, Laos, Oman, and the United States, all of which missing from the original data.

The original results are not robust to using 2022 data: in Model 4, the coefficient for *Trade/GDP* is negative and not statistically significant. This is because PPP indicators were revised after 2014 to adopt newer ICP benchmarks. Consequently, "PPP data are now provided only from 1990, as the longer the time period between the estimate and the

Table 3: The Effect of Trade Dependence on Genuine Savings (Random Effects GLS), Including Only Observations Available From All Sources, 1980–1999

	(1) Original Model	(2) WDI 2002	(3) WDI 2012	(4) WDI 2022
Trade/GDP (ln)	2.053	1.912	1.779	−0.181
	(1.304)	(1.303)	(1.397)	(3.097)
GNI pc (ln)	25.009**	24.102**	24.598***	51.776
	(10.070)	(10.057)	(9.232)	(37.989)
GNI pc^2 (ln)	−1.190**	−1.160*	−1.164**	−3.395
	(0.594)	(0.593)	(0.534)	(2.178)
Economic Growth	0.016	0.020	0.027	0.131
	(0.042)	(0.042)	(0.049)	(0.086)
Agriculture/GDP	−0.080	−0.130*	−0.001	0.138
	(0.073)	(0.074)	(0.091)	(0.192)
Currency Crisis	0.421	0.458	1.253*	1.726
	(0.594)	(0.594)	(0.740)	(1.226)
Fuel Exporter	−21.870***	−22.094***	−20.993***	−4.186
	(3.161)	(3.158)	(3.395)	(25.065)
Democracy	0.746	0.703	−0.855	0.241
	(1.030)	(1.028)	(1.146)	(2.534)
Political Constraints	−0.942	−0.949	−3.491*	−4.927
	(1.847)	(1.844)	(2.059)	(4.208)
Government Stability	−0.627	−0.621	−0.677	0.195
	(0.483)	(0.483)	(0.619)	(0.998)
Population Density (ln)	1.814***	1.834***	1.379**	1.751
	(0.556)	(0.555)	(0.598)	(4.190)
Population Size (ln)	−0.059	−0.116	0.731	−1.254
	(0.559)	(0.559)	(0.608)	(3.767)
Population Urban	−8.244***	−8.635***	−7.373***	14.016
	(2.234)	(2.232)	(2.612)	(13.373)
Civil War	−0.474	−0.491	−0.085	1.214
	(0.899)	(0.898)	(1.068)	(1.927)
Peace Years	0.028	0.028	0.013	0.035
	(0.035)	(0.035)	(0.037)	(0.107)
Constant	−96.740**	−87.473**	−107.309***	−231.806
	(42.707)	(42.747)	(38.624)	(164.192)
Number of Observations	735	735	735	735
Countries	95	95	95	95

This is a replication of Model 1, Table 1 in de Soysa and Neumayer (2005). Standard errors appear in parentheses. All regressions assume an AR1 correlation structure and include year dummies. All independent variables are lagged one year.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

benchmark, the greater the risk of inaccuracy" (World Bank, 2023). If the authors used 2022 data, they would not find support for their argument — not because this argument is wrong,

but because the sample covered by each vintage affects one's empirical conclusions. The number of observations shrinks from 2,069 in Model 1 to 840 in Model 4. These observations are not evenly lost across all countries, as Figure 2 shows. Recent vintages might be closer to the "truth," but if "truthful" values are not available for all countries, the sample — and the empirical results — will be biased.

The discrepancies identified in Table 2 are not just a function of sample selection bias, but also a consequence of data revisions: different vintages might disagree about the same country-year pairs. Table 3 reduces the analysis to the 735 country-year pairs common to all vintages. When all vintages have the exact same coverage, the effect of trade on genuine savings, while never significant, is positive in Models 1 to 3 and negative in Model 4.

4.2 Vreeland (2008)

Many studies only derive their control variables from the WDI, not their main variables. This is still empirically consequential: since different vintages cover different countries and years, vintaged control variables affect the sample size. Moreover, coefficients tend to be biased downward when there are measurement errors on the right-hand side, attenuating the "true" effect that would appear had variables been measured correctly (Hausman, 2001). Researchers might be *underestimating* the substantive importance of their findings, as the second replication shows.

Using data on 109 dictatorships between 1985 and 1996, Vreeland (2008) finds that multiparty dictatorships are more likely to torture opponents *and* more likely to enter the United Nations Convention Against Torture (CAT) than one-party or no-party dictatorships. When power is shared, there is more room to disagree with the ruling party. Since at least some dissent is tolerated, defection is more common, as is the punishment of defectors. But interest groups can force the regime to make concessions — and entering the CAT is one concession. The focus on dictatorships is valuable: autocrats overstate their growth rates and do not revise these figures (Martínez, 2022), so there could be fewer discrepancies between

WDI releases than in the previous replication.

Table 4: The Effect of Multiple Parties on Torture in Dictatorships (Ordinal Logit), 1985–1996

	(1) Original Model	(2) WDI 1998	(3) WDI 2004	(4) WDI 2018
Parties	0.578***	0.447***	0.507***	−0.066
	(0.149)	(0.152)	(0.148)	(0.246)
GDP/Capita	0.016	0.044*	−0.006	0.003
	(0.025)	(0.026)	(0.025)	(0.011)
Growth	0.007**	−0.003	0.002	−0.001
	(0.003)	(0.015)	(0.013)	(0.012)
Population	0.002***	0.003***	0.003***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)
Trade/GDP	−0.010***	−0.009***	−0.011***	−0.010***
	(0.002)	(0.002)	(0.002)	(0.002)
Civil War	0.795***	1.072***	0.856***	0.576**
	(0.170)	(0.184)	(0.176)	(0.241)
Communist	−1.098***	−1.655***	−1.561***	−2.968***
	(0.355)	(0.396)	(0.346)	(0.618)
Cut 1	−3.073***	−2.891***	−3.069***	−3.720***
	(0.241)	(0.244)	(0.255)	(0.380)
Cut 2	−1.048***	−0.910***	−1.163***	−1.533***
	(0.165)	(0.177)	(0.202)	(0.274)
Cut 3	1.141***	1.312***	0.960***	0.315
	(0.172)	(0.186)	(0.198)	(0.268)
Cut 4	2.700***	2.942***	2.505***	1.716***
	(0.219)	(0.239)	(0.236)	(0.301)
Number of Observations	694	668	710	403
Log Likelihood	−893.6	−852.1	−927.6	−535.3

This is a replication of Model 1, Table 1 in Vreeland (2008). Robust standard errors appear in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

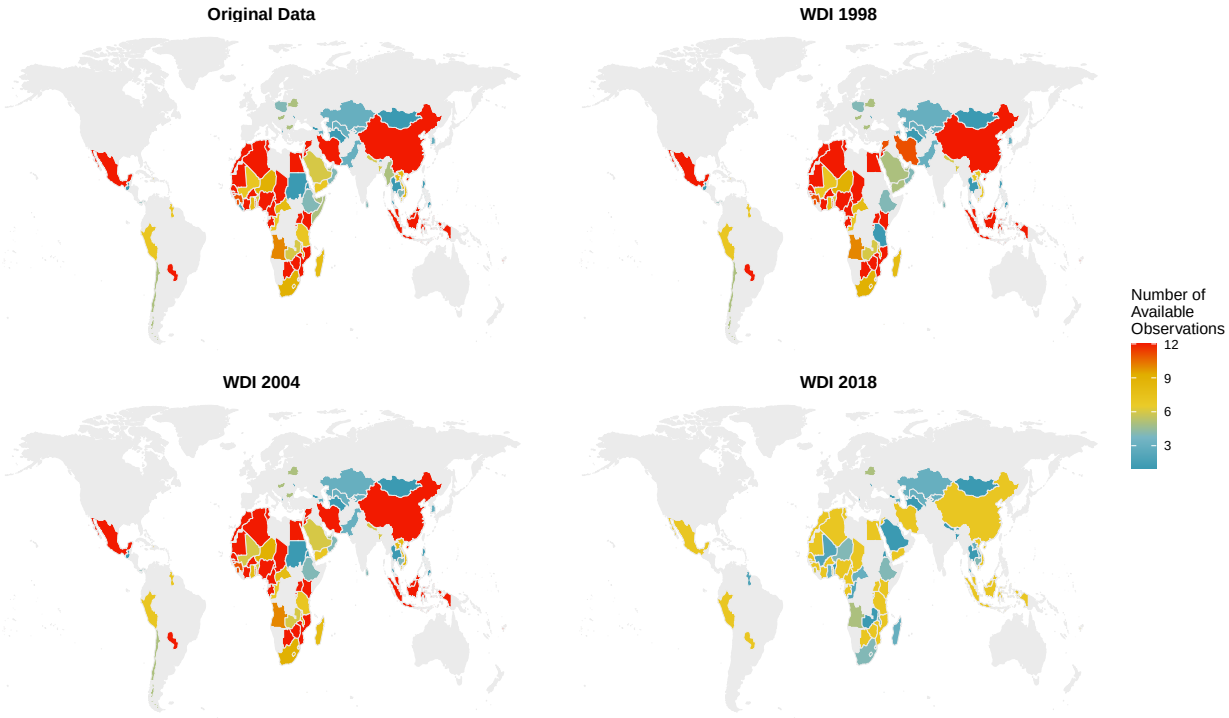
I replicate the first part of Vreeland's argument: multiparty dictatorships are more likely to engage in torture. The outcome is a five-point ordinal scale of torture, ranging from one (no allegations of torture) to five (torture is prevalent or widespread), and the estimated model is an ordinal logit. The main explanatory variable, *Parties*, takes the value of one if more than one party exists legally, and zero otherwise. Four control variables come from the WDI: GDP per capita in 1995 PPP dollars, GDP growth (%), population, and trade (% of GDP). The study also controls for communist regimes and the occurrence of a civil war.

Vreeland uses 2004 WDI data, combined with PWT 6.1. I compare the original data

to three WDI releases: 1998 (the earliest release for which all required years are available), 2004 (without PWT additions), and 2018. These vintages report GDP per capita with 1987, 1995, and 2011 as the base year, respectively.

Table 4 presents the results of this replication. Rather than interpret the coefficients for control variables, I investigate how their inclusion affects the main results. Controlling for GDP per capita, GDP growth, population, and trade using 1998 or 2004 data, *Parties* continues to have a significant positive effect on the outcome. The coefficient for *Parties* is smaller in Model 2 than in Model 3, confirming that older benchmarks are biased downward: their measurement errors attenuate the "true" effect of multiple parties on torture.

Figure 3: Available Observations by Country, 1985–1996



These maps show the number of complete observations available by country for the variables included in Vreeland's analysis. Since Vreeland examines 109 separate dictatorships during the 1985–1996 period, the maps only report the availability of data for these country-years.

Model 4 indicates that the original results are not robust to using 2018 data. The number of observations shrinks from 694 (Model 1) to 403 (Model 4) because the World Bank ceased

to provide data before 1990 after revising PPP indicators in 2014. This should not be taken as evidence against the original findings: since the CAT was opened for signature in 1984, the reduced sample drops crucial years and only covers a fraction of the 109 dictatorships included in the original study. Figure 3 reiterates that researchers using different vintages would draw inferences from different samples, as the missing observations are not the same. Even if only the control variables are vintaged, their inclusion might shape our substantive conclusions about the relationship between other unvintaged variables.

Table 5: The Effect of Multiple Parties on Torture in Dictatorships (Ordinal Logit), Including Only Observations Available From All Sources, 1985–1996

	(1) Original Model	(2) WDI 1998	(3) WDI 2004	(4) WDI 2018
Parties	0.250	0.284	0.268	0.332
	(0.314)	(0.316)	(0.316)	(0.315)
GDP/Capita	0.078	0.081	0.094	0.070***
	(0.051)	(0.068)	(0.061)	(0.025)
Growth	0.006	−0.003	−0.010	−0.012
	(0.018)	(0.026)	(0.025)	(0.026)
Population	0.006***	0.006***	0.006***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)
Trade/GDP	−0.015***	−0.018***	−0.015***	−0.018***
	(0.004)	(0.004)	(0.004)	(0.004)
Civil War	0.327	0.321	0.298	0.319
	(0.284)	(0.286)	(0.280)	(0.288)
Communist	−3.279***	−3.301***	−3.221***	−3.026***
	(0.813)	(0.831)	(0.828)	(0.830)
Cut 1	−4.072***	−4.354***	−4.086***	−4.135***
	(0.545)	(0.562)	(0.545)	(0.536)
Cut 2	−1.446***	−1.714***	−1.458***	−1.506***
	(0.375)	(0.397)	(0.375)	(0.377)
Cut 3	0.629*	0.390	0.618*	0.597
	(0.362)	(0.379)	(0.363)	(0.372)
Cut 4	2.214***	1.998***	2.206***	2.208***
	(0.399)	(0.417)	(0.401)	(0.413)
Number of Observations	283	283	283	283
Log Likelihood	−354.405	−351.280	−353.962	−351.121

This is a replication of Model 1, Table 1 in Vreeland (2008). Robust standard errors appear in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5 reduces the analysis to the 283 country-year pairs common to all vintages. When all vintages have the same coverage, all models agree that multiparty dictatorships are more

likely to torture opponents but disagree about the magnitude of the effect — a discrepancy driven by data revisions, not just by changes in the sample size. This underscores the need to look beyond point estimates, which might suffer from measurement error.

4.3 Goldstein, Rivers, and Tomz (2007)

The final replication looks at variation across different data sources, not just different releases of the same source. According to Rose (2004), formal membership in the General Agreement on Tariffs and Trade (GATT) and the World Trade Organization (WTO) did little to increase trade. However, Goldstein et al. (2007) argue that the GATT/WTO created rights and obligations even for countries that had not attained formal membership, like colonies and newly independent states. Using dyadic data from 1946 to 2004, the authors introduce a measure of participation that goes beyond formal members to include nonmember participants; using this measure, formal members and nonmember participants trade more than nonparticipants.

I do not replicate Goldstein et al.'s main results, but rather a gravity model used to establish Rose's original finding, without the novel measure of GATT/WTO participation. The outcome is the value of imports (in 1967 USD) from country i to country j. The key explanatory variables indicate the existence of a unilateral or bilateral GATT/WTO membership. Besides the standard gravity variables (the distance between i and j as well as the product of their GDP), the model controls for participation in preferential trade agreements (PTA) or in the Generalized System of Preferences (GSP), currency unions, land area, colonial ties, shared language or border, and whether the two countries are islands or landlocked.

Combining data from multiple sources into one single variable appears to be standard practice to maximize coverage. This study is no different. The main source for *Log Product Real GDP* (in 1967 USD) is the 2005 WDI, complemented by version 6.1 of the PWT and the 2003 Maddison Project, which report GDP in 2000 USD, 1996 USD, and 1990 international

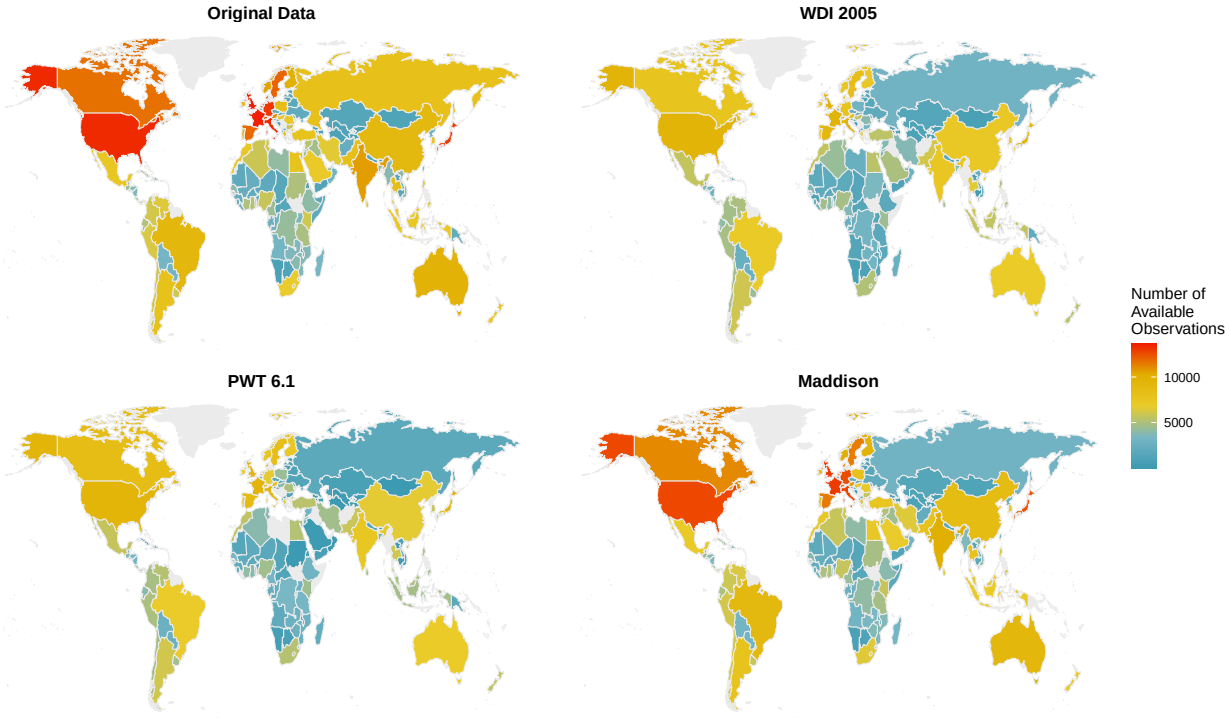
Table 6: The Effect of GATT/WTO Membership on Trade (Ordinary Least Squares), 1946–2004

	(1) Original Model	(2) WDI 2005	(3) PWT 6.1	(4) Maddison 2003
Both Formal GATT/WTO Members	−0.070***	−0.182***	0.120***	0.107***
	(0.026)	(0.031)	(0.031)	(0.027)
Only One Formal GATT/WTO Member	−0.211***	−0.289***	−0.223***	−0.203***
	(0.025)	(0.032)	(0.032)	(0.027)
Reciprocal PTA	0.334***	0.197***	0.289***	0.287***
	(0.027)	(0.030)	(0.035)	(0.030)
Nonreciprocal PTA	0.139***	0.178***	0.181***	0.160***
	(0.035)	(0.036)	(0.038)	(0.037)
GSP	−0.097***	−0.086***	0.214***	0.177***
	(0.022)	(0.023)	(0.027)	(0.025)
Currency Union	1.010***	1.134***	1.069***	1.129***
	(0.075)	(0.086)	(0.093)	(0.079)
Colonial Orbit	1.755***	1.771***	1.639***	1.548***
	(0.104)	(0.258)	(0.153)	(0.100)
Log Product Real GDP	0.771***	0.752***	0.907***	0.844***
	(0.005)	(0.005)	(0.007)	(0.006)
Log of Distance	−0.708***	−0.830***	−0.795***	−0.754***
	(0.015)	(0.017)	(0.019)	(0.016)
Common Language	0.357***	0.269***	0.469***	0.421***
	(0.034)	(0.038)	(0.042)	(0.038)
Land Border	0.577***	0.562***	0.471***	0.438***
	(0.059)	(0.068)	(0.080)	(0.066)
Number of Landlocked	−0.142***	−0.124***	−0.197***	−0.218***
	(0.020)	(0.022)	(0.025)	(0.022)
Number of Islands	0.237***	0.288***	0.231***	0.224***
	(0.032)	(0.035)	(0.037)	(0.034)
Log Product Land Area	−0.095***	−0.043***	−0.161***	−0.134***
	(0.005)	(0.005)	(0.006)	(0.005)
Constant	−11.754***	−13.221***	−18.447***	−17.085***
	(0.252)	(0.278)	(0.358)	(0.321)
Number of Observations	381,656	269,313	243,109	360,730
R^2	0.613	0.643	0.631	0.583

This is a replication of Model 1, Table 1 in Goldstein et al. (2007). Robust standard errors, clustered by directed dyad, appear in parentheses. All regressions include year dummies.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

dollars, respectively. It is not clear how these data were rescaled to 1967 USD, as the Maddison Project does not provide nominal GDP data that would enable such calculations. Therefore, I use WDI, PWT, and Maddison data in their raw form, without changing the

Figure 4: Available Observations by Country, 1946–2004



These maps show the number of complete observations available, by country, for the variables included in Goldstein et al.'s analysis. Since the authors use dyadic data, each observation is counted twice, once for every member of a dyad. Maddison coverage begins in 1945, before the decolonization of Africa, so there are more observations available for developed than for developing countries. WDI coverage begins in 1960 and PWT coverage begins in 1950.

base year.

Instead of interpreting control variable coefficients, I ask whether their inclusion affects the main results. According to Model 1 in Table 6, formal GATT/WTO membership significantly *reduces* trade. Model 2, estimated using WDI data, corroborates this finding (which is unsurprising; the original data rely primarily on the WDI). Models 3 and 4 find the opposite: trade *increases* when both members of the dyad are formal GATT/WTO members.[4] Using PWT or Maddison data, the authors would find support for their argument even without the updated participation measure.[5]

[4] Linsi and Mügge (2019) also show that Rose's findings are not consistent across mirror trade statistics: when both members of the dyad are formal GATT/WTO members, imports *decrease*, but exports *increase*.

[5] Goldstein et al.'s main results, using the updated participation measure, are robust to data changes (see appendix).

Table 7: The Effect of GATT/WTO Membership on Trade, Including Only Observations Available From All Sources (Ordinary Least Squares), 1946–2004

	(1) Original Model	(2) WDI 2005	(3) PWT 6.1	(4) Maddison 2003
Both Formal GATT/WTO Members	−0.133***	−0.110***	0.111***	0.088**
	(0.033)	(0.034)	(0.035)	(0.036)
Only One Formal GATT/WTO Member	−0.259***	−0.292***	−0.267***	−0.376***
	(0.034)	(0.036)	(0.037)	(0.038)
Reciprocal PTA	0.229***	0.147***	0.239***	0.227***
	(0.032)	(0.033)	(0.037)	(0.036)
Nonreciprocal PTA	0.099***	0.176***	0.138***	0.138***
	(0.037)	(0.038)	(0.039)	(0.039)
GSP	−0.083***	−0.060**	0.239***	0.224***
	(0.024)	(0.025)	(0.028)	(0.028)
Currency Union	1.055***	1.222***	1.048***	1.167***
	(0.097)	(0.102)	(0.104)	(0.102)
Colonial Orbit	1.321***	1.505***	1.310***	1.374***
	(0.277)	(0.250)	(0.257)	(0.238)
Log Product Real GDP	0.829***	0.763***	0.919***	0.900***
	(0.006)	(0.006)	(0.008)	(0.008)
Log of Distance	−0.792***	−0.837***	−0.830***	−0.834***
	(0.018)	(0.019)	(0.020)	(0.020)
Common Language	0.434***	0.349***	0.486***	0.443***
	(0.040)	(0.042)	(0.045)	(0.045)
Land Border	0.604***	0.568***	0.543***	0.522***
	(0.075)	(0.076)	(0.083)	(0.085)
Number of Landlocked	−0.153***	−0.124***	−0.235***	−0.153***
	(0.024)	(0.025)	(0.026)	(0.027)
Number of Islands	0.230***	0.239***	0.232***	0.211***
	(0.035)	(0.037)	(0.039)	(0.039)
Log Product Land Area	−0.108***	−0.050***	−0.170***	−0.156***
	(0.006)	(0.006)	(0.006)	(0.006)
Constant	−13.350***	−13.539***	−18.862***	−18.751***
	(0.290)	(0.305)	(0.373)	(0.375)
Number of Observations	202,819	202,819	202,819	202,819
R^2	0.682	0.668	0.643	0.640

This is a replication of Model 1, Table 1 in Goldstein et al. (2007). Robust standard errors, clustered by directed dyad, appear in parentheses. All regressions include year dummies.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 4 confirms that different sources cover different periods, but this alone cannot explain the discrepancies between results. Table 7 shows that these discrepancies persist when the analysis includes only the 202,819 country-year pairs common to all three sources.

This is because measures like real GDP, GDP growth, and trade to GDP rely on nominal GDP information that differs across sources due to currency conversions or PPP adjustments. Compared to the PWT, the WDI consistently overestimates the size of developed economies and underestimates the economy of small nations. Just as researchers using different vintages might draw inferences from different samples, the choice of one source over another can affect researchers' conclusions.

5 Conclusions

Vintaged data are ubiquitous in political science. However, 31.21% of the studies published in *International Organization* from 2005 to 2020 do not mention their vintage, while 13.87% do not mention their source. Besides sharing their replication code and data, researchers should disclose their sources and vintages, as de Soysa and Neumayer (2005), Vreeland (2008), and Goldstein et al. (2007) do.

Even rigorous findings might disappear if re-estimated using different data. To identify potential sources of imprecision, researchers can consult the World Bank's Data Updates and Errata (2023), which describe "additions, deletions, and changes in codes, descriptions, definitions, sources and topics." Plotting the distribution of variables can help identify extreme values. A plot of GDP growth over time would reveal that Equatorial Guinea's economy grew 18.2% in 2000. Using case knowledge, researchers can assess whether this is a "true" outlier — the result of increased oil production — or a product of human error — in which case Equatorial Guinea would appear in the Data Updates and Errata (it does not).

Researchers should strike a balance between maximizing coverage and using the most recent data. Unless working with pre-1950 data, one should favor recent PWT or WDI releases, which are revised more frequently than Maddison data. But there are trade-offs: more recent series might have worse coverage or cover an entirely different sample, as the replication of Vreeland (2008) illustrates. Depending on the sample of interest, older releases

are preferable, even if further away from the "truth."

Finally, researchers should avoid mixing different vintages or sources. If a vintage or source suffers from inherent measurement error, this error must be consistent across all observations. Off-the-shelf datasets are very convenient, but their use does not absolve researchers from thinking about the quality of their data and, ultimately, the robustness of their findings.

References

Alt, J., Lassen, D. D. and Wehner, J. (2014), 'It Isn't Just about Greece: Domestic Politics, Transparency and Fiscal Gimmickry in Europe', *British Journal of Political Science* **44**(4), 707–716.

Aragão, R. and Linsi, L. (2022), 'Many Shades of Wrong: What Governments Do When They Manipulate Statistics', *Review of International Political Economy* **29**(1), 88–113.

Berry, F., Iommi, M., Stanger, M. and Venter, L. (2018), 'The Status of GDP Compilation Practices in 189 Economies and the Relevance for Policy Analysis', *IMF Working Paper* **37**.

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

Ciccone, A. and Jarociński, M. (2010), 'Determinants of Economic Growth: Will Data Tell?', *American Economic Journal: Macroeconomics* **2**(4), 222–246.

Croushore, D. and Stark, T. (2003), 'A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?', *Review of Economics and Statistics* **85**(3), 605–617.

de Soysa, I. and Neumayer, E. (2005), 'False Prophet, or Genuine Savior? Assessing the

Effects of Economic Openness on Sustainable Development, 1980–99', *International Organization* **59**(3), 731–772.

Deaton, A. and Aten, B. (2017), 'Trying to Understand the PPPs in ICP 2011: Why Are the Results So Different?', *American Economic Journal: Macroeconomics* **9**(1), 243–64.

Fariss, C. J., Anders, T., Markowitz, J. N. and Barnum, M. (2022), 'New Estimates of Over 500 Years of Historic GDP and Population Data', *Journal of Conflict Resolution* **66**(3), 553–591.

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

Gleditsch, K. S. (2002), 'Expanded Trade and GDP Data', *Journal of Conflict Resolution* **46**(5), 712–724.

Goldstein, J. L., Rivers, D. and Tomz, M. (2007), 'Institutions in International Relations: Understanding the Effects of the GATT and the WTO on World Trade', *International Organization* **61**(1), 37–67.

Hausman, J. (2001), 'Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left', *Journal of Economic Perspectives* **15**(4), 57–67.

Hollyer, J. R., Rosendorff, B. P. and Vreeland, J. R. (2014), 'Measuring Transparency', *Political Analysis* **22**(4), 413–434.

Inklaar, R. and Prasada Rao, D. S. (2017), 'Cross-Country Income Levels over Time: Did the Developing World Suddenly Become Much Richer?', *American Economic Journal: Macroeconomics* **9**(1), 265–90.

Johnson, S., Larson, W., Papageorgiou, C. and Subramanian, A. (2013), 'Is Newer Better? Penn World Table Revisions and Their Impact on Growth Estimates', *Journal of Monetary Economics* **60**(2), 255–274.

Kerner, A., Jerven, M. and Beatty, A. (2017), 'Does It Pay to Be Poor? Testing for Systematically Underreported GNI Estimates', *Review of International Organizations* **12**(1), 1–38.

Linsi, L. and Mügge, D. K. (2019), 'Globalization and the Growing Defects of International Economic Statistics', *Review of International Political Economy* **26**(3), 361–383.

Mansfield, E. D. and Reinhardt, E. (2008), 'International Institutions and the Volatility of International Trade', *International Organization* **62**(4), 621–652.

Martínez, L. R. (2022), 'How Much Should We Trust the Dictator's GDP Growth Estimates?', *Journal of Political Economy* **130**(10), 2731–2769.

Pinkovskiy, M. and Sala-i Martin, X. (2020), 'Shining a Light on Purchasing Power Parities', *American Economic Journal: Macroeconomics* **12**(4), 71–108.

Ram, R. and Ural, S. (2014), 'Comparison of GDP Per Capita Data in Penn World Table and World Development Indicators', *Social Indicators Research* **116**(2), 639–646.

Rose, A. K. (2004), 'Do We Really Know That the WTO Increases Trade?', *The American Economic Review* **94**(2), 98–114.

Vreeland, J. R. (2008), 'Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture', *International Organization* **62**(1), 65–101.

World Bank (2023), *Data Updates and Errata*.

URL: *https://datahelpdesk.worldbank.org/knowledgebase/articles/906522-data-updates-and-errata*

Appendix

A De Soysa and Neumayer (2005)

In the main text, Tables 2 and 3 replicate a random effects generalized least squares model corresponding to Model 1, Table 1 of the original study (de Soysa and Neumayer, 2005, 750-751). In this model, economic globalization is measured as trade in % of GDP, using data from the WDI. But he original study also employs a second measure of globalization, foreign direct investment (FDI) stock in % of GDP, using data from the United Nations.

Tables A.1 and A.2 replicate Models 2 and 3 in Table 1 of the original study, respectively; both models include this second measure of globalization. As these tables show, the authors would have found mixed results had they used the 2012 or 2022 version of the WDI. In Models 3 and 4 of Table A.1, the two measures of economic globalization — *Trade/GDP* and *FDI/GDP* — have opposing and contradicting effects.

Table A.1: The Effect of FDI Dependence on Genuine Savings (Random Effects GLS), 1980–1999

	(1) Original Model	(2) WDI 2002	(3) WDI 2012	(4) WDI 2022
FDI/GDP (ln)	0.516*	0.685*	−0.436	2.121*
	(0.286)	(0.371)	(0.386)	(1.159)
GNI pc (ln)	21.999***	20.923***	14.711**	26.494
	(6.311)	(8.044)	(6.430)	(33.774)
(GNI pc)^2 (ln)	−0.964***	−0.963**	−0.694*	−1.977
	(0.374)	(0.480)	(0.377)	(1.919)
Economic Growth	−0.011	0.030	0.109***	0.097
	(0.024)	(0.031)	(0.031)	(0.079)
Agriculture/GDP	−0.060	−0.140**	−0.203***	0.072
	(0.046)	(0.059)	(0.062)	(0.173)
Currency Crisis	0.234	−0.078	1.350**	0.449
	(0.395)	(0.516)	(0.558)	(1.083)
Fuel Exporter	−17.853***	−22.212***	−21.421***	−2.088
	(2.446)	(2.749)	(2.627)	(20.435)
Democracy	1.035	0.940	0.691	1.185
	(0.717)	(0.946)	(0.937)	(2.153)
Political Constraints	−1.189	1.184	−2.509	−4.729
	(1.500)	(1.982)	(1.957)	(3.700)
Government Stability	−0.387	−0.385	−0.178	0.190
	(0.344)	(0.454)	(0.467)	(0.905)
Population Density (ln)	1.215***	1.533***	1.382***	1.939
	(0.439)	(0.499)	(0.491)	(3.647)
Population Size (ln)	−0.332	−0.445	−0.201	−0.307
	(0.385)	(0.438)	(0.427)	(3.191)
Population Urban	−9.095***	−8.524***	−6.669***	15.705
	(1.542)	(1.843)	(1.805)	(11.986)
Civil War	−1.702**	1.205	−0.019	1.003
	(0.741)	(0.967)	(0.953)	(1.787)
Peace Years	0.002	0.011	0.010	0.034
	(0.026)	(0.033)	(0.032)	(0.091)
Constant	−73.212***	−66.653**	−36.458	−149.759
	(26.164)	(33.351)	(26.243)	(146.247)
Observations	2,050	2,041	1,817	840
Countries	135	136	122	109

This is a replication of Model 2, Table 1 in de Soysa and Neumayer (2005). Standard errors appear in parentheses. All regressions assume an AR1 correlation structure and include year dummies. All independent variables are lagged one year.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: The Effect of Trade and FDI Dependence on Genuine Savings (Random Effects GLS), 1980–1999

	(1) Original Model	(2) WDI 2002	(3) WDI 2012	(4) WDI 2022
Trade/GDP (ln)	2.122***	1.692*	3.277***	−1.519
	(0.784)	(1.022)	(1.074)	(2.762)
FDI/GDP (ln)	0.382	0.544	−0.689*	2.196*
	(0.290)	(0.379)	(0.392)	(1.168)
GNI pc (ln)	20.211***	19.343**	13.736**	27.369
	(6.306)	(8.059)	(6.372)	(33.798)
(GNI pc)^2 (ln)	−0.858**	−0.869*	−0.625*	−2.017
	(0.374)	(0.481)	(0.374)	(1.919)
Economic Growth	−0.008	0.030	0.110***	0.096
	(0.024)	(0.031)	(0.031)	(0.079)
Agriculture/GDP	−0.046	−0.129**	−0.172***	0.068
	(0.047)	(0.061)	(0.063)	(0.174)
Currency Crisis	0.003	−0.281	1.182**	0.568
	(0.398)	(0.523)	(0.561)	(1.106)
Fuel Exporter	−18.037***	−22.337***	−21.570***	−2.045
	(2.426)	(2.739)	(2.581)	(20.349)
Democracy	1.103	1.042	0.752	1.184
	(0.715)	(0.947)	(0.934)	(2.155)
Political Constraints	−1.057	1.309	−2.282	−4.731
	(1.496)	(1.982)	(1.954)	(3.704)
Government Stability	−0.358	−0.358	−0.157	0.170
	(0.344)	(0.455)	(0.468)	(0.907)
Population Density (ln)	1.052**	1.405***	1.100**	1.997
	(0.440)	(0.505)	(0.492)	(3.634)
Population Size (ln)	0.064	−0.136	0.495	−0.621
	(0.410)	(0.476)	(0.478)	(3.226)
Population Urban	−9.007***	−8.381***	−6.910***	15.408
	(1.534)	(1.841)	(1.779)	(11.970)
Civil War	−1.667**	1.236	−0.049	1.036
	(0.739)	(0.967)	(0.952)	(1.791)
Peace Years	0.001	0.011	0.010	0.034
	(0.026)	(0.033)	(0.031)	(0.091)
Constant	−80.629***	−72.193**	−56.033**	−142.177
	(26.432)	(33.922)	(26.701)	(146.810)
Observations	2,046	2,035	1,814	840
Countries	135	135	122	109

This is a replication of Model 3, Table 1 in de Soysa and Neumayer (2005). Standard errors appear in parentheses. All regressions assume an AR1 correlation structure and include year dummies. All independent variables are lagged one year.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Vreeland (2008)

In the main text, Table 4 replicates Model 1, Table 1 of the original study (Vreeland, 2008, 83), corresponding to the first part of the author's argument: multiparty dictatorships are more likely to engage in torture. As a reminder, the outcome variable is a five-point ordinal scale of torture ranging from one (no allegations of torture) to five (torture is prevalent or widespread).

Below, I replicate the remaining models in Table 1 of the original study (Vreeland, 2008, 83): a fixed effects logit (Table B.1) and a duration dependence logit (Table B.2). *Communist* drops out of the specifications in Table B.2 because it does not vary by country in this sample.

Table B.1: The Effect of Multiple Parties on Torture in Dictatorships (Fixed Effects Logit), 1985–1996

	(1) Original Model	(2) WDI 1998	(3) WDI 2004	(4) WDI 2018
Parties	0.715**	0.748**	0.752**	0.569
	(0.344)	(0.367)	(0.360)	(0.650)
GDP/Capita	−0.329	−0.290	−0.644*	−0.208
	(0.348)	(0.501)	(0.380)	(0.227)
Growth	0.007	−0.036*	−0.027	−0.021
	(0.014)	(0.020)	(0.019)	(0.018)
Population	0.116***	0.122***	0.145***	−0.080
	(0.042)	(0.042)	(0.045)	(0.117)
Trade/GDP	−0.006	−0.010	−0.015*	0.007
	(0.010)	(0.009)	(0.009)	(0.009)
War	0.570	0.540	0.536	−1.008
	(0.466)	(0.470)	(0.460)	(0.771)
Number of Observations	428	401	427	217
Log Likelihood	−162.029	−152.001	−158.845	−79.006

This is a replication of Model 2, Table 1 in Vreeland (2008). Robust standard errors appear in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: The Effect of Multiple Parties on Torture in Dictatorships (Duration Dependence Logit), 1985–1996

	(1) Original Model	(2) WDI 1998	(3) WDI 2004	(4) WDI 2018
Parties	0.798***	0.626***	0.690***	−0.150
	(0.217)	(0.229)	(0.225)	(0.316)
GDP/Capita	−0.014	−0.000	0.004	−0.006
	(0.032)	(0.040)	(0.031)	(0.015)
Growth	0.016**	−0.029**	−0.018	−0.015
	(0.007)	(0.014)	(0.013)	(0.011)
Population	0.001*	0.002*	0.002**	0.016**
	(0.001)	(0.001)	(0.001)	(0.007)
Trade/GDP	−0.009***	−0.009***	−0.008**	−0.006
	(0.003)	(0.003)	(0.003)	(0.004)
Civil War	0.407*	0.727***	0.522**	0.194
	(0.241)	(0.238)	(0.243)	(0.328)
Communist	−0.688	−0.916	−1.035	−1.814
	(0.679)	(0.856)	(0.847)	(1.275)
Count	−1.289***	−1.197***	−1.282***	−1.321***
	(0.203)	(0.209)	(0.201)	(0.288)
Spline 1	−0.006	−0.003	−0.003	0.001
	(0.010)	(0.009)	(0.008)	(0.008)
Spline 2	−0.131***	−0.115***	−0.117***	−0.100**
	(0.036)	(0.037)	(0.035)	(0.044)
Spline 3	0.086**	0.068	0.067*	0.043
	(0.042)	(0.043)	(0.038)	(0.042)
Constant	−0.059	−0.158	−0.098	0.678*
	(0.275)	(0.294)	(0.295)	(0.397)
Number of Observations	694	668	710	403
Log Likelihood	−310.987	−285.382	−311.291	−180.569

This is a replication of Model 3, Table 1 in Vreeland (2008). Robust standard errors appear in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Finally, I replicate the second part of the argument: multiparty dictatorships are more likely to enter the CAT. The main empirical evidence supporting this argument is provided by Table 3 (Vreeland, 2008, 90), which reports five different Weibull hazard models with different dependent variables and specifications. Table B.3 replicates Models 2 and 4 in Vreeland's Table 3. I only replicate these two fully specified Weibull hazard models because these are the only models that include the WDI control variables and whose results could potentially change from one data vintage to another.

Table B.3: The Effect of Multiple Political Parties on CAT Participation in Dictatorships (Weibull Hazard Models), 1985–1996

	Signing — Hazard Ratios Reported				Ratification — Hazard Ratios Reported			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Original Model	WDI 1998	WDI 2004	WDI 2018	Original Model	WDI 1998	WDI 2004	WDI 2018
Parties	2.875**	2.947***	2.978***	6.821	2.180	2.089	2.867**	6.827
	(1.221)	(1.156)	(1.257)	(13.580)	(1.110)	(1.000)	(1.535)	(13.198)
Log Torture	1.192	1.359	1.520	9.653	0.895	0.987	1.394	5.882
	(0.504)	(0.653)	(0.694)	(14.347)	(0.458)	(0.551)	(0.727)	(7.751)
Communist	2.652*	3.790**	2.057	0.000***	1.128	2.224	1.998	0.000***
	(1.542)	(2.190)	(1.517)	(0.000)	(1.250)	(2.166)	(1.999)	(0.000)
Regional Score	0.639	1.122	0.742	4.724	9.964**	26.157***	7.965**	23.303
	(0.872)	(1.995)	(0.934)	(15.620)	(11.102)	(29.781)	(7.289)	(114.303)
Number Under	0.960	0.924	0.953	1.005	0.970	0.939	0.963	0.992
	(0.040)	(0.053)	(0.036)	(0.031)	(0.035)	(0.043)	(0.025)	(0.037)
Muslim	1.899	1.698	1.693	1.174	1.561	1.627	1.208	3.899
	(1.054)	(0.933)	(0.875)	(1.373)	(0.868)	(0.958)	(0.682)	(4.793)
GDP/Capita	1.062	0.970	1.025	1.056	1.067	0.933	1.036	1.039
	(0.047)	(0.073)	(0.052)	(0.048)	(0.054)	(0.071)	(0.064)	(0.044)
Population	1.001	1.000	1.001	1.003	1.002	1.001	1.001	0.981
	(0.001)	(0.000)	(0.001)	(0.025)	(0.001)	(0.001)	(0.001)	(0.014)
Trade/GDP	0.994	1.000	1.002	0.996	0.995	1.001	1.002	0.998
	(0.004)	(0.005)	(0.006)	(0.008)	(0.005)	(0.005)	(0.007)	(0.007)
p	1.572	1.895	1.510	0.961	1.536	2.124	1.541	1.242
Number of Observations	483	482	514	250	558	556	588	295

This is a replication of Models 2 and 4, Table 3 in Vreeland (2008). Robust standard errors appear in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In Table B.3, the dependent variables indicate the time until a dictatorship *signs* the CAT and the time until a dictatorship *ratifies* the CAT, respectively. The reported estimates are hazard ratios (that is, exponentiated coefficients). As before, this replication shows that the original results are robust to using GDP per capita, GDP growth, population, and trade data from different WDI vintages, with the exception of the 2018 WDI vintage, for which 20% of all observations are missing.

C Goldstein, Rivers, and Tomz (2007)

To construct their GDP variable, Goldstein et al. (2007, 50) "turned first to the 2005 edition of the World Bank's *World Development Indicators* ... then extended the series backwards to 1946, using U.S. dollar figures from the Penn World Tables, the United Nations, the Oxford Latin American Economic History Database, and the IMF *International Financial Statistics*. In a few cases, we used the GDP indices from Maddison ... to complete the data set." Table 6 replicates the original analysis using *only* the 2005 WDI (Model 2), *only* version 6.1 of the PWT (Model 3), and *only* the 2003 Maddison Project (Model 4). In an additional replication, I combine different sources — as Goldstein et al. do — but use newer vintages of each source.

For this additional replication, I collect GDP data from the most recent version of each source as of February 2023: the 2022 WDI (in 2010 USD, 1960–2004); version 10.01 of the PWT (in 2017 USD, 1950–2004); the UN Data website (in 2015 USD, 1970–2004); the Montevideo-Oxford Latin American Economic History Database, or MOxLAD (in 1970 USD, 1946–2004); the International Financial Statistics (in domestic currency, with variable base years, 1950–2004); and the 2020 Maddison database (in 2011 USD, 1946–2004). Since each source reports a different base year, Goldstein et al. convert all data to 1967 USD using nominal GDP values or a GDP deflator (that is, the ratio of nominal to real GDP for 1967).

I convert the updated data to 2010 USD, rather than 1967 USD, for two reasons. First, this is the base year used by the new vintage of the main source, the 2022 WDI. Second, UN Data coverage as of February 2023 only begins in 1970; UN Data could be converted to 1967 USD using the WDI deflator, but this would lead to a loss of all observations that are present in the UN data and not in the WDI.

My replication has two limitations. First, the Maddison Project provides neither nominal GDP data nor a GDP deflator. I use the WDI deflator to make these conversions, but this leads to a loss of some observations (though this is not a major problem because the authors only use Maddison data "[i]n a few cases"). Second, the International Financial

Table C.1: The Effect of GATT/WTO Membership on Trade (Ordinary Least Squares), 1946–2004

	(1) Original Model	(2) WDI 2022, PWT 10.01, and Others
Both Formal GATT/WTO Members	−0.070***	−0.076***
	(0.026)	(0.028)
Only One Formal GATT/WTO Member	−0.211***	−0.158***
	(0.025)	(0.028)
Reciprocal PTA	0.334***	0.217***
	(0.027)	(0.028)
Nonreciprocal PTA	0.139***	0.114***
	(0.035)	(0.036)
GSP	−0.097***	−0.066***
	(0.022)	(0.023)
Currency Union	1.010***	0.875***
	(0.075)	(0.078)
Colonial Orbit	1.755***	1.481***
	(0.104)	(0.174)
Log Product Real GDP	0.771***	0.744***
	(0.005)	(0.005)
Log of Distance	−0.708***	−0.772***
	(0.015)	(0.016)
Common Language	0.357***	0.258***
	(0.034)	(0.037)
Land Border	0.577***	0.542***
	(0.059)	(0.067)
Number of Landlocked	−0.142***	−0.165***
	(0.020)	(0.021)
Number of Islands	0.237***	0.251***
	(0.032)	(0.033)
Log Product Land Area	−0.095***	−0.076***
	(0.005)	(0.005)
Constant	−11.754***	−13.144***
	(0.252)	(0.279)
Number of Observations	381,656	333,281
R^2	0.613	0.593

This is a replication of Model 1, Table 1 in Goldstein et al. (2007). Robust standard errors, clustered by directed dyad, appear in parentheses. All regressions include year dummies.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Statistics only provide data in domestic currency and with variable base years; since there is no straightforward way to convert this to one single currency and one single base year, I disregard this data source.

In Table C.1, Model 1 is the original model estimated by Goldstein et al. As described above, Model 2 uses newer vintages of Goldstein et al.'s data sources, converting these data to 2010 USD. The results are remarkably similar, though Model 2 has considerably fewer observations, reflecting the fact that newer vintages of the same source tend to discard older observations because "the longer the time period between the estimate and the benchmark, the greater the risk of inaccuracy" (World Bank, 2023).

Table C.2: The Effect of GATT/WTO Membership on Trade, Including Dyad and Year Effects (Ordinary Least Squares), 1946–2004

	(1) Original Model	(2) WDI 2005	(3) PWT 6.1	(4) Maddison
Both Formal GATT/WTO Members	0.070***	−0.044*	0.009	0.037*
	(0.020)	(0.023)	(0.023)	(0.020)
Only One Formal GATT/WTO Member	−0.016	−0.178***	−0.127***	−0.116***
	(0.019)	(0.024)	(0.025)	(0.020)
Reciprocal PTA	0.348***	0.272***	0.332***	0.313***
	(0.022)	(0.025)	(0.026)	(0.023)
Nonreciprocal PTA	−0.070**	−0.074**	0.059*	0.017
	(0.032)	(0.033)	(0.033)	(0.032)
GSP	−0.105***	−0.015	−0.016	0.001
	(0.019)	(0.020)	(0.022)	(0.020)
Currency Union	0.491***	0.451***	0.562***	0.569***
	(0.088)	(0.078)	(0.121)	(0.093)
Colonial Orbit	0.880***	0.505**	0.610***	0.798***
	(0.081)	(0.220)	(0.137)	(0.090)
Log Product Real GDP	0.665***	0.913***	0.998***	1.035***
	(0.011)	(0.025)	(0.023)	(0.019)
Constant	−15.291***	−29.544***	−35.638***	−37.054***
	(0.542)	(1.270)	(1.165)	(0.975)
Number of Observations	381,656	269,313	243,109	360,730
R^2	0.839	0.872	0.867	0.839

This is a replication of Model 2, Table 1 in Goldstein et al. (2007). Robust standard errors, clustered by directed dyad, appear in parentheses. All regressions include year and dyad dummies.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

While Tables 6, 7, and C.1 replicate Model 1, Table 1 of Goldstein et al. (2007, 53), which includes year dummies, Table C.2 replicates Model 2, Table 1 of the original study, which includes dyad dummies in addition to year dummies. In the original analysis, GATT participation no longer has a clear effect on trade once dyad dummies are included: the

coefficients for *Both Formal GATT/WTO Members* and *Only One Formal GATT/WTO Member* have opposite signs. Had Goldstein et al. used only WDI data (Model 2 in Table C.2), they would have obtained more consistent results, as both variables have negative coefficients that are statistically significant.

Finally, Table C.3 replicates the main results (Models 1 and 2, Table 3 in Goldstein et al. 2007, 54). All models in Table C.3 confirm the authors' expectation that formal members and nonmember participants of the GATT/WTO trade more than nonparticipants, though Models 2 and 6 (using only WDI data) show that this effect is not statistically significant when only one of the countries in the dyad participates in the GATT/WTO regime as a formal member.

Table C.3: The Effect of GATT/WTO Membership on Trade, Using Different Membership/Participation Measures (Ordinary Least Squares), 1946–2004

	Full Model				Restricted Model			
	(1) Original Model	(2) WDI 2005	(3) PWT 6.1	(4) Maddison	(5) Original Model	(6) WDI 2005	(7) PWT 6.1	(8) Maddison
Both Participate in the GATT/WTO					0.354***	0.246***	0.284***	0.348***
					(0.034)	(0.047)	(0.052)	(0.035)
Both Formal Members	0.341***	0.205***	0.259***	0.316***				
	(0.035)	(0.047)	(0.053)	(0.036)				
Both Nonmember Participants	0.447***	0.472***	0.358***	0.547***				
	(0.070)	(0.105)	(0.089)	(0.076)				
Formal Member and Nonmember Participant	0.381***	0.309***	0.301***	0.382***				
	(0.037)	(0.051)	(0.055)	(0.039)				
Only One Participates in the GATT/WTO					0.200***	0.068	0.117**	0.132***
					(0.029)	(0.043)	(0.048)	(0.031)
Formal Member	0.200***	0.043	0.090*	0.112***				
	(0.030)	(0.044)	(0.049)	(0.032)				
Nonmember Participant	0.173***	0.210***	0.257***	0.216***				
	(0.040)	(0.063)	(0.063)	(0.043)				
Reciprocal PTA	0.344***	0.272***	0.331***	0.309***	0.343***	0.270***	0.332***	0.309***
	(0.022)	(0.025)	(0.026)	(0.023)	(0.022)	(0.025)	(0.026)	(0.023)
Nonreciprocal PTA	−0.052	−0.065**	0.066**	0.033	−0.053*	−0.066**	0.063*	0.031
	(0.032)	(0.033)	(0.033)	(0.032)	(0.032)	(0.033)	(0.033)	(0.032)
GSP	−0.098***	−0.013	−0.010	0.006	−0.099***	−0.009	−0.006	0.010
	(0.019)	(0.019)	(0.022)	(0.020)	(0.019)	(0.020)	(0.022)	(0.020)
Currency Union	0.496***	0.462***	0.565***	0.581***	0.492***	0.452***	0.554***	0.566***
	(0.088)	(0.078)	(0.121)	(0.093)	(0.088)	(0.078)	(0.120)	(0.092)
Colonial Orbit	0.808***	0.433**	0.565***	0.694***	0.836***	0.487**	0.593***	0.766***
	(0.082)	(0.219)	(0.139)	(0.089)	(0.081)	(0.220)	(0.136)	(0.089)
Log Product Real GDP	0.661***	0.910***	0.991***	1.032***	0.661***	0.899***	0.985***	1.025***
	(0.011)	(0.025)	(0.023)	(0.019)	(0.011)	(0.025)	(0.023)	(0.019)
Constant	−15.379***	−29.642***	−35.527***	−37.135***	−15.377***	−29.137***	−35.239***	−36.840***
	(0.541)	(1.269)	(1.161)	(0.972)	(0.542)	(1.263)	(1.164)	(0.972)
Number of Observations	381,656	269,313	243,109	360,730	381,656	269,313	243,109	360,730
R^2	0.840	0.872	0.868	0.840	0.840	0.872	0.868	0.840

This is a replication of Models 1 and 2, Table 2 in Goldstein et al. (2007). Robust standard errors, clustered by directed dyad, appear in parentheses. All regressions include year and dyad dummies. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.