Labeling Laggards and Leaders: International Organizations and the Politics of Defining Development

Supplementary Information

September 13, 2018

Contents

List of Tables

List of Figures

S1 Expert Survey

While my theory predicts which classifications influence which actors, it does not actually predict the direction of the effects. To derive directional predictions would demand developing additional theories for the costs and benefits each actor faces for allocating aid or investment to one country over another, or rating a country a certain way. While the directions of my hypotheses are intuitive and consistent with previous findings, there are sometimes plausible reasons to believe effects could occur in the opposite direction. In fact, Knack et al. (2014) begins with competing hypotheses regarding donor behavior: Donors could add to the countries they observe other donors funding, or they could do the opposite in an attempt to compensate for the behavior of others.

Prior to conducting my analysis, I conducted a brief survey of experts at the Center for Global Development (CGD) to solicit their priors not only on where the effects would be but also on the direction. CGD is one of the most actively involved think tanks in IDA graduation policy reform and many of the fellows have decades of experience in multilateral organizations working to promote development through aid, investment, finance, and governance initiatives.[1] To sufficiently define my research question for them, I presented respondents with a list of the outcomes I would be examining, potential explanations for any effects, and potential explanations for null effects.

The results appear in Table S1. First, the surveyed experts expressed mostly uniform beliefs about the direction of any potential effect of crossing a threshold. The group tended to believe that threshold crossings would decrease aid, increase investment, and improve credit ratings, democracy ratings, and the incumbent's chances of being re-elected. This provides me with a helpful basis for judging whether any of my effects are surprising or counter-intuitive.

Second, there was substantial variation in beliefs about which classifications would produce effects, and which outcomes would be affected. Taken together, my respondents thought classification effects were more likely to exist for aid and FDI than for ratings and re-election probabilities. They also thought classification effects were more likely to exist on the operational category change (graduating from IDA) rather than the analytical one (crossing the LIC ceiling).[2] But even within each outcome and each classification system,

[1] While all staff were invited to participate in the survey, I subset my sample to only my 17 respondents who reported 3 or above on a 5-point scale when asked about their expertise/familiarity with IDA graduation policy.

[2] This survey question is not exactly comparable to my results, since I asked experts about the category change from IDA-only or Blend to IBRD-only status, not about crossing the operational cutoff. Nonetheless, it is a good proxy for the perceived importance of operational categories.

there was significant variation among experts in their beliefs about which classifications would matter and whose behaviors they would affect. At the very least, this underscores that the expected effects of classification are not obvious and that better theories and evidence are needed to understand this phenomenon.

Table S1: Survey of expert priors from CGD

Effect of:	Crossing LIC ceiling			Graduating from IDA		
Expected direction of effect:	-	+	null	-	+	null
Aid	10 (.59)	1 (.06)	6 (.35)	14 (.82)	0	3 (.18)
FDI	0	12 (.71)	5 (.29)	1 (.06)	13 (.76)	3 (.18)
Credit rating	0	8 (.47)	9 (.53)	0	12 (.71)	5 (.29)
Other ratings (e.g. democracy)	0	4 (.24)	13 (.76)	0	5 (.29)	12 (.71)
Prob. of incumbent re-election	0	4 (.24)	13 (.76)	1 (.06)	6 (.35)	10 (.59)

Note: N=17. Table reports count, with frequency in parentheses.

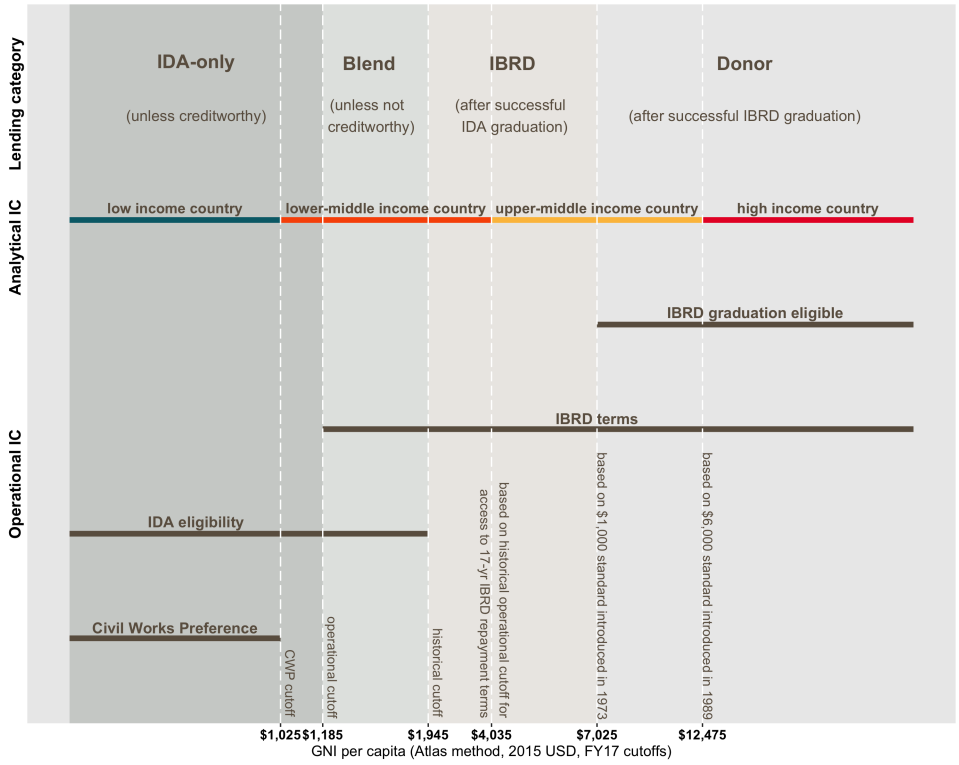
S2 Details about Classification Systems

Although they are commonly conflated, the World Bank actually produces two income classifications: analytical and operational classifications. Both classifications sort countries into four groups by calculating whether the country's gross national income (GNI) per capita, a measure of a country's overall income divided by its population, falls above or below certain pre-defined thresholds separating categories.[3] Although the two systems share some thresholds, other thresholds differ, and the two systems are used for different purposes. Only operational classifications are used for determining whether a country will be able to access finance from IDA, IBRD, or both. In contrast, the analytical classifications are produced by the World Bank for analytical purposes but have no bearing on World Bank policy. It is therefore a widespread misconception that "low-income countries" (an analytical grouping) are synonymous with IDA recipients, when in fact there are more IDA-eligible countries that are "lower-middle income" than "low-income."

The analytical classification system labels countries as "low income," "lower-middle income," "upper-

[3]The measure that determines a country's classification in both systems is GNI per capita, valued annually in US dollars using the Atlas method. The Atlas method uses an average of the current year's exchange rate and the previous two years' exchange rates inflated to current year prices and is employed to reduce volatility of the estimation of GNI per capita. To calculate exchange rates, the Atlas method compares a country's prices to a weighted average of prices from the economies that have Special Drawing Rights in the IMF. This is known as an "SDR deflator." Skeptics of this system contend that the current measure is too blunt to meaningfully capture development, while its defenders argue that it is the only indicator consistently and accurately reported by nearly all countries. The more indicators and adjustments incorporated into the calculation or classification of income, the less universal and comparable the results will be. For example, one levied critique is that using GNI data without adjusting for purchasing power parity (PPP, or how much a dollar buys in each economy) fails to take into account important variation in price levels between countries. This concern has been raised repeatedly both to World Bank data technocrats and the World Bank's Board of Executive Directors, but both groups have each time concluded that the coverage and quality of price data is insufficient to produce a robust cross-national data set of income. See Fantom and Serajuddin (2016).

Figure S1: Income classifications and lending categorizations in the World Bank



middle income," or "high income" (LIC/LMIC/UMIC/HIC). These classifications are presented in color in the middle of Figure S1, and can be easily calculated by determining where a country's GNI per capita is relative to a series of thresholds. Every year on July 1, the World Bank Development Economics Group publishes its new classifications, which are made on the basis of GNI data from the previous calendar year and on the historical thresholds, which are updated only to account for inflation. These classifications were never intended to be and have never been used for operational purposes.

For operational decisions, the World Bank uses a different set of classifications, shown in black lines in the bottom half of Figure S1. These classifications are an important but not the sole determinant of whether countries may borrow from IDA, IBRD, or both. The most important operational cutoff for my study is the one that triggers the multi-year process of graduating from IDA, and it is simply referred to as the "operational cutoff." When a country crosses this cutoff, the country desk will request an assessment of the country's creditworthiness from the IBRD to determine whether the country will be able to access IBRD loans. If it is found creditworthy, it becomes reclassified as a "Blend" country for the following fiscal year and can access both IDA and IBRD loans.[4] After a country has crossed the operational cutoff for three

[4]However, the terms on its old IDA credits will harden. Also, if it is not found creditworthy, it becomes a "gap" country: the

consecutive years and has been found creditworthy by the IBRD, IDA reviews its case for graduation during the next replenishment meeting, a meeting that occurs every three years in which donors pledge to replenish the stock of IDA that will be lent to countries in the next cycle.

As is apparent from this discussion, income is an important criterion in the type of assistance a country receives from the World Bank, but it is not the only criterion. Creditworthiness, for instance, also plays an important role in a country's ability to borrow from IBRD. Other criteria appear when a country is evaluated for graduating from IDA. Because graduation from IDA is very costly for states, and because the institution wants to avoid reverse graduations, IDA staff or deputies frequently recommend against graduation if they anticipate political or economic instability for any number of reasons.[5] The result of these negotiations is that, in practice, countries spend an average of six years in Blend status (i.e. borrowing from both IDA and IBRD) before they are graduated entirely from IDA.

Many have investigated the historical origins of these thresholds and have found that they were arbitrarily selected for reasons that are irrelevant to today's world.[6] Recall from the introduction that the analytical classifications were born of the need for a simple analytical instrument to track development progress. When selecting thresholds to formalize this system in 1989, researchers identified several figures from World Bank policies that had previously existed but were already defunct. The LIC ceiling, distinguishing LICs from LMICs, is actually based on a cutoff introduced in the 1970s called the "civil works preference" cutoff, below which countries received preferences in civil works procurement bids in Bank-financed projects because they were not thought to be competitive enough. The threshold separating LMICs from UMICs was based on a different operational threshold no longer in use: The cutoff used by the IBRD to assess 15-year versus 17-year repayment terms, categories that have since been collapsed.[7] As for operational policy, the most important cutoff has always been the one determining when countries would begin the process of graduating from IDA. The first operational cutoff was introduced in 1964 at $250 per capita, but by 1989, IDA's resources were too limited to accommodate all countries falling below this cutoff. IDA lowered the opera-

country may still access IDA on hardened terms, but cannot yet access IBRD.

[5]IDA graduation is costly because countries lose access to future IDA credits and often face accelerated repayments on old IDA credits. In addition, borrowing limits in IBRD sometimes mean that the total volume of borrowing will be constrained. In the most recent graduation policy review document, IDA staff supported its recommendation for graduating three of the twelve current Blend countries by citing the following indicators: poverty headcount ratio, Human Development Index, real output growth, nominal public debt, commodity exports, political stability and absence of violence/terrorism, risk of debt distress, Worldwide Governance Indicators, and Economic Vulnerability Index. In case descriptions, the authors mention low prices on commodity exports and lack of market access. World Bank IDA Resource Mobilization Department (2016)

[6]See Nielsen (2011); Knack et al. (2014); Fantom and Serajuddin (2016).

[7]While I don't focus on it in this study, the UMIC/HIC threshold was simply set at $6,000, and no rationale has ever been found for this number.

tional cutoff by 38%, which was $580 per capita in US dollars. The old operational cutoff became known as the "historical cutoff." Although it has no operational significance anymore, IDA graduation reviews often note the number of years a country has exceeded the historical cutoff when recommending graduation. As such, the historical cutoff is informally relevant for IDA-eligibility. In other words, when states cross the "operational cutoff," this triggers the process of graduation, and when states cross the "historical cutoff," this makes the need for graduating them more urgent. Both of these operational cutoffs, however, are not related to any current operational policy, and if any strategy contributed to their selection, it does not systematically separate countries any longer.

While the origin of each threshold is different, their shared characteristic is that they are all historical artifacts that make them exogenous to today's development environment. As such, while a country's GNI per capita is not random, whether its GNI per capita falls on one side of a threshold or another is. This makes it possible to devise an empirical strategy that will identify the effect of the classification itself when we control for a country's income. In this study, I will use the LIC ceiling and the LMIC ceiling to study the effects of the analytical classification system. I omit some higher cutoffs—namely, the UMIC ceiling and the IBRD graduation threshold—to preserve a focus on how classification affects developing countries. For other works exploiting World Bank eligibility cutoffs, see Carnegie and Samii (2017); Buntaine et al. (2017).

S3 Summary Statistics

Table S2: Table of summary statistics

Group	Mean	SD	Min	Max	N
GNI per capita					
All	3619.85	8253.43	60.00	203900.00	3867
L	409.59	198.32	60.00	1020.00	1290
LM	1799.29	817.30	430.00	4050.00	1423
UM	5812.16	2388.77	1700.00	12550.00	878
H	21036.67	23462.91	6260.00	203900.00	276
ODA Disbursements (All donors)					
All	395.02	667.59	-959.96	11428.02	3846
L	651.80	863.64	11.84	11428.02	1290
LM	379.36	576.99	-943.15	5509.01	1423
UM	151.30	344.10	-959.96	3441.78	877
H	23.11	92.22	-460.26	559.30	256
FDI (percentage of GDP)					
All	3.95	7.31	-101.37	161.83	3703
L	3.32	6.91	-7.29	90.46	1238
LM	3.50	5.22	-32.29	62.30	1351
UM	4.98	5.47	-39.69	36.88	861
H	5.93	17.07	-101.37	161.83	253
Creditworthiness (IIR)					
All	33.63	17.68	4.95	84.60	2370
L	20.47	9.48	4.95	63.10	791
LM	31.84	12.32	7.20	76.20	888
UM	47.03	14.90	14.85	82.35	556
H	67.30	11.95	23.40	84.60	135
Freedom House PR score (flipped)					
All	4.33	2.02	1.00	7.00	3723
L	3.22	1.68	1.00	7.00	1285
LM	4.50	1.86	1.00	7.00	1350
UM	5.31	1.87	1.00	7.00	858
H	5.83	2.08	1.00	7.00	230
Logged population					
All	15.49	2.10	9.19	21.03	3864
L	16.07	1.62	11.58	20.94	1287
LM	15.40	2.17	10.85	21.01	1423
UM	15.06	2.39	9.19	21.03	878
H	14.59	1.99	10.30	18.78	276
Logged gross capital formation					
All	21.46	2.91	0.00	29.20	3519
L	20.06	3.49	0.00	26.66	1206
LM	21.74	2.15	0.00	28.51	1234
UM	22.53	2.30	17.44	29.20	832
H	23.21	1.88	19.01	26.93	247

S4 Heterogeneous Aid Effects

An additional implication of my theory is that some donors will be more susceptible to the strategic mechanism than others and should therefore exhibit stronger classification effects. For example, Honig (2019) illustrates that while some agencies receive considerable control over their operations, others are highly restricted by unpredictable and contingent sources of funding. These latter "insecure agencies," which are extensively monitored and evaluated, are more likely to prioritize the appearance of success over prudence in development interventions. Given their institutional design, I would expect these non-autonomous agencies to rely more heavily on classifications, since they must perpetually justify their decisions and behaviors.[8] Classifications can also provide a useful way for aid agencies to signal their excellent performance to audiences, such as domestic constituencies or even to their peers.[9] When aid agencies face scrutiny from either of these audiences about the impartiality of their decisions, they will be more likely to use classifications.

To investigate this additional hypothesis, I explore heterogeneity in the classification effect by donor. I find that classifications have a stronger hold over donors with a greater need to signal their commitments to development. As a first cut, I separate the results by bilateral and multilateral donors. Multilateral institutions are widely perceived as more impartial than bilateral institutions, so only bilateral institutions should have a need for classifications as a signaling device.[10] As expected, Table S3 finds that this result is primarily driven by bilateral donors; multilateral assistance is not significantly affected by any country category. Interestingly, the effects are even stronger for the non-traditional donors than they are for traditional donors, even though non-traditional donors are widely thought to ignore need in their foreign aid policy.[11] One interpretation of this finding could be that non-traditional donors have a reputation for using aid as political instruments, and so they have greater need of classifications to overcome this stigma. However, this conclusion should be interpreted with caution, as my data only reflects the aid given by non-traditional donors who report their statistics to the OECD, a non-random sample of potentially more transparent non-traditional donors.[12]

[8] Similarly, democratic donors have greater strategic incentives to use classifications in their decision-making, as they must justify these behaviors to domestic audiences.

[9] Honig (2019) argues that aid agencies try to improve their performance on the Aid Transparency Index out of concern for their social reputation in the donor community. It is possible that aid agencies wish to show each other, not their funders, that they benefit the neediest. This possibility still follows the strategic logic in the sense that donors use classifications to signal to audiences.

[10] See Milner (2006).

[11] "Traditional" donors are those that are members of the OECD's DAC.

[12] These include: Azerbaijan, Bulgaria, Croatia, Cyprus, Estonia, Israel, Kazakhstan, Kuwait, Latvia, Liechtenstein, Lithuania, Malta, Romania, Russia, Saudi Arabia, Chinese Taipei, Thailand, Timor-Leste, Turkey, and the UAE.

Table S3: The effects of classifications on aid by type of donor

	(1) DAC	(2) Non-DAC	(3) Multilateral
	A. Above LIC ceiling		
Above LIC ceiling (t-1)	−0.036	0.141	−0.035
	(0.069)	(0.106)	(0.078)
Constant	0.548	12.545*	2.040
	(5.364)	(7.073)	(4.852)
Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Period F.E.	✓	✓	✓
Period	3-Year	3-Year	3-Year
Observations	631	585	629
R^2	0.882	0.630	0.879
	B. Above LMIC ceiling		
Above LMIC ceiling (t-1)	−0.339**	−0.426***	−0.165
	(0.133)	(0.154)	(0.103)
Constant	2.340	14.984**	2.940
	(4.808)	(6.831)	(4.784)
Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Period F.E.	✓	✓	✓
Period	3-Year	3-Year	3-Year
Observations	631	585	629
R^2	0.884	0.633	0.880

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: The table reports coefficients from OLS regressions of the outcome on a dummy variable coded 1 if a country is above the cutoff, controlling for GNI per capita. Standard errors are clustered at the country level. Covariates include lagged values of log population, log gross capital formation, and Freedom House political rights score. The sample is restricted to countries that have ever benefited from IDA after 1987. All dependent variables have been standardized for ease of comparison.

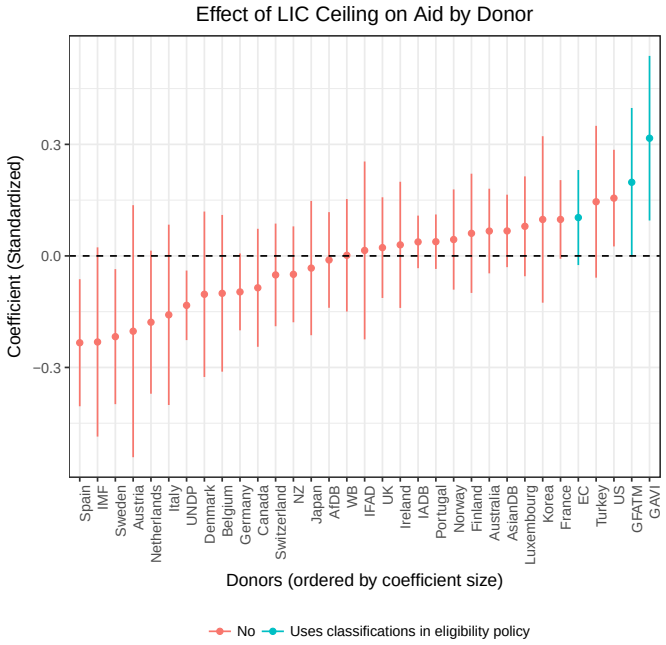
I further examine the importance of the strategic mechanism in Figure S2. I estimate the classification effect separately by each individual donor and plot them in order of effect sizes. Since I observed earlier that aid was most strongly affected by the LMIC ceiling, I am primarily interested in these effects, which appear in Panel B. Unsurprisingly, the strongest negative effects are for Gavi and the Global Fund, two donors who formally include the World Bank classifications in their eligibility policies.[13] These cases are useful because they are far more likely to be explained by the strategic than the cognitive mechanism: Institutionalizing an eligibility policy requires great deliberation and justification, tempering cognitive biases but amplifying any

[13] While they exhibit the strongest effects, these donors provide a relatively small amount of overall aid, so they do not drive the overall finding in the main results.

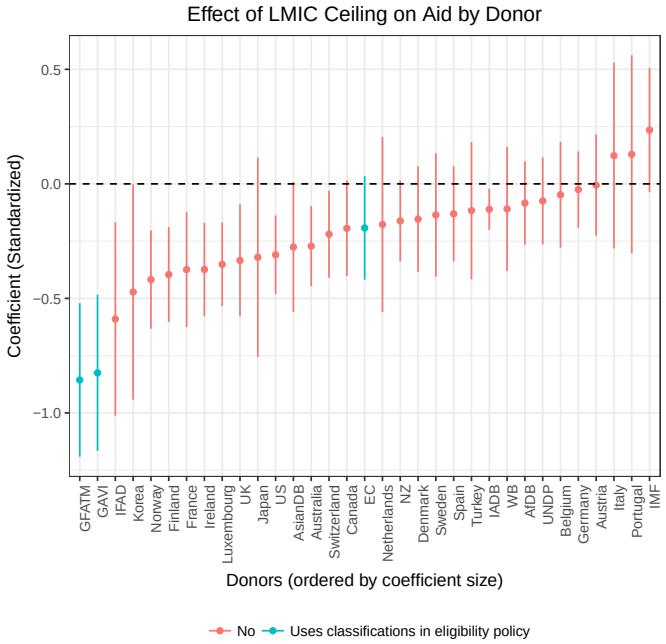
audience effects. What is especially interesting is that these same cases actually do respond to the LIC ceiling as well (Panel A), but the effect is opposite to my original predictions. These strategically-motivated donors actually reward countries for becoming LMICs but punish them when they become UMICs. These results also partially suggest that the original null finding of LIC graduation on aid may obscure important heterogeneity: For example, Germany and Sweden do give less aid to countries once they become LICs. In Panel B, we see many donors, such as Scandinavian donors, exhibiting the classification effect. While these findings should be subjected to multiple comparisons corrections before drawing inferences about any individual donor's behavior, these results suggest important heterogeneity in the responses of donors and their susceptibility to the two mechanisms.

Figure S2: The effects of classifications by individual donors

(a) LIC Ceiling



(b) LMIC Ceiling



Note: Each observation represents the effect of classifications on the official development assistance of a single donor, noted on the x-axis. Each observation comes from a different regression and observations are ordered by coefficient size. Donors in blue (Gavi, the Global Fund, and the European Community) include the World Bank's classifications as part of their eligibility policies.

S5 Robustness of Main Model

Table S4: Robustness to alternative measure of FDI

	(1) FDI inflows (levels)	(2) FDI inflows (perc. of GDP)
	A. Above LIC ceiling	
Above LIC ceiling (t-1)	0.011	0.022
	(0.069)	(0.142)
Constant	−0.121	5.217
	(4.078)	6.018
Covariates	✓	✓
Country F.E.	✓	✓
Period F.E.	✓	✓
Period	Year	Year
Observations	3,062	3,110
R^2	0.833	.340
	B. Above LMIC ceiling	
Above LMIC ceiling (t-1)	−0.032	−0.082
	(0.048)	(0.074)
Constant	0.032	5.579
	(4.123)	(5.933)
Covariates	✓	✓
Country F.E.	✓	✓
Period F.E.	✓	✓
Period	Year	Year
Observations	3,062	3,110
R^2	0.833	0.340

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: The table reports coefficients from OLS regressions of the outcome on a dummy variable coded 1 if a country is above the cutoff, controlling for GNI per capita. Standard errors are clustered at the country level. Covariates include lagged values of log population, log gross capital formation, and Freedom House political rights score. The Freedom House political rights score is inverted so that positive values are more democratic. All dependent variables have been standardized for ease of comparison.

Table S5: Robustness to dropping Freedom House control

	(1) Aid	(2) FDI inflows	(3) Creditworthiness
	A. Above LIC ceiling		
Above LIC ceiling (t-1)	−0.013	0.014	.069
	(0.066)	(0.067)	(0.059)
Constant	2.733	−0.475	−3.244
	(5.308)	(4.027)	(3.864)
F.H. Covariate	∅	∅	∅
Other Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Period F.E.	✓	✓	✓
Period	3-Year	Year	Year
Observations	645	3,137	3,018
R^2	0.884	0.834	0.851
	B. Above LMIC ceiling		
Above LMIC ceiling (t-1)	−0.233**	−0.024	0.124*
	(0.109)	(0.047)	(0.064)
Constant	3.891	−0.395	−4.367
	(5.086)	(4.068)	(3.986)
F.H. Covariate	∅	∅	∅
Other Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Period F.E.	✓	✓	✓
Period	3-Year	Year	Year
Observations	645	3,137	3,018
R^2	0.885	0.834	0.851

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: The table reports coefficients from OLS regressions of the outcome on a dummy variable coded 1 if a country is above the cutoff, controlling for GNI per capita. Standard errors are clustered at the country level. Covariates include lagged values of log population and log gross capital formation. All dependent variables have been standardized for ease of comparison.

Table S6: Robustness to dropping all covariates

	(1) Aid	(2) FDI inflows	(3) Creditworthiness	(4) Democracy
		A. Above LIC ceiling		
Above LIC ceiling (*t*-1)	−0.009	−0.010	0.061	0.127**
	(0.063)	(0.062)	(0.054)	(0.064)
Constant	2.698***	−3.424***	−2.076***	−1.291***
	(0.429)	(0.236)	(0.251)	(0.296)
Covariates	∅	∅	∅	∅
Country F.E.	✓	✓	✓	✓
Period F.E.	✓	✓	✓	✓
Period	3-Year	Year	Year	Year
Observations	712	3,400	3,279	3,487
R^2	0.892	0.811	0.842	0.828
		B. Above LMIC ceiling		
Above LMIC ceiling (*t*-1)	−0.261***	0.012	0.129**	0.034
	(0.089)	(0.048)	(0.063)	(0.066)
Constant	2.526***	−3.390***	−1.987***	−1.450***
	(0.342)	(0.256)	(0.254)	(0.303)
Covariates	∅	∅	∅	∅
Country F.E.	✓	✓	✓	✓
Period F.E.	✓	✓	✓	✓
Period	3-Year	Year	Year	Year
Observations	712	3,400	3,279	3,487
R^2	0.894	0.811	0.843	0.828

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: The table reports coefficients from OLS regressions of the outcome on a dummy variable coded 1 if a country is above the cutoff, controlling for GNI per capita. Standard errors are clustered at the country level. In the aid regressions, the sample is restricted to countries that have ever benefited from IDA after 1987. All dependent variables have been standardized for ease of comparison.

Table S7: Robustness to yearly observations of aid

	(1) Aid	(2) Aid	(3) Aid
	A. Above LIC ceiling		
Above LIC ceiling (t-1)	−0.022	−0.019	
	(0.063)	(0.055)	
Above LIC ceiling (t-2)			0.013
			(0.049)
Constant	3.151	4.914	1.537
	(5.124)	(4.211)	(3.735)
Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Year F.E.	✓	✓	✓
Period	3-Year Period	Year	Year
Observations	632	1,852	1,764
R^2	0.886	0.852	0.859
	B. Above LMIC ceiling		
Above LMIC ceiling (t-1)	−0.246**	−0.176**	
	(0.107)	(0.087)	
Above LMIC ceiling (t-2)			−0.134
			(0.088)
Constant	4.374	5.982	2.328
	(4.874)	(4.080)	(3.703)
Covariates	✓	✓	✓
Country F.E.	✓	✓	✓
Year F.E.	✓	✓	✓
Period	3-Year Period	Year	Year
Observations	632	1,852	1,764
R^2	0.887	0.853	0.859

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: The table reports coefficients from OLS regressions of the outcome on a dummy variable coded 1 if a country is above the cutoff, controlling for GNI per capita. Standard errors are clustered at the country level. Covariates include lagged values of log population, log gross capital formation, and Freedom House political rights score. All dependent variables have been standardized for ease of comparison.

S6 Alternative Model: Graduations

The baseline specification uses levels and holds GNI per capita constant to identify the effect of being above a cutoff; an alternative approach is to use change scores and hold changes in GNI per capita constant to identify the effect of crossing a cutoff. This specification focuses on explaining year-to-year variation in how observers react to a single country's change in category, omitting the country-to-country variation in how observers react to countries just above and just below a threshold.

To operationalize this, I take as my dependent variable the standardized difference between the outcome in year *t* and in year *t*-1. I regress this change score on two dummy variables indicating whether the country crossed the threshold from below (graduated) or above (reverse graduated) since the previous year, and also the change in GNI per capita since the previous year. The main estimands of interest are the coefficient on the graduation and reverse graduation terms. Consistent with the main results, changes in income and classification are lagged by a year, so they take place between year *t*-2 and year *t*-1 to produce changes in the outcome variable between year *t*-1 and year *t*. I continue to include year fixed effects but remove country fixed effects since change scores account for this.

$$\begin{aligned} Y_{i,t} - Y_{i,t-1} = \alpha + \beta_1 Graduated_{i,t-1} + \beta_2 Reverse\ graduated_{i,t-1} \\ + \delta(log(GNIpc)_{i,t-1} - log(GNIpc)_{i,t-2}) + \tau_t + \varepsilon \end{aligned} \tag{1}$$

The results appearing in Table S8 suggest interesting similarities and differences with those reported in the baseline specification. Consistent with the previous results, no threshold crossing produces observable changes in net inflows of FDI (column 2). Also consistent with the previous results, graduating from LMIC to UMIC significantly improves perceptions of the country's creditworthiness, and the effect is larger and more significant than in the baseline specification (column 3). This suggests that the variation we observed in the baseline specification is primarily driven by credit raters reacting to news of a country's graduation as a positive signal, rather than comparing countries side-by-side. Column 4 shows results that differ from the baseline model but are also consistent with Hypothesis 3: In these results, crossing the LMIC threshold but not the LIC threshold influences democracy scores. In contrast with the baseline model, neither graduation produces a reaction from donors. This suggests that the previous findings are better explained by donors allocating scarce resources in favor of lower-income countries, rather than reacting to changes in country

Table S8: Effects of graduations and reverse graduations

	(1) Aid	(2) FDI inflows	(3) Creditworthiness	(4) Democracy
		A. LIC to LMIC		
Graduated	−0.048	−0.035	−0.056	−0.019
	(0.104)	(0.172)	(0.169)	(0.090)
Reverse graduated	0.140	0.185	0.098	−0.216
	(0.180)	(0.200)	(0.248)	(0.222)
Log GNIpc change	−0.142	0.024	1.320***	0.061
	(0.148)	(0.137)	(0.180)	(0.161)
Constant	−0.202***	0.030	−0.317***	−0.135
	(0.065)	(0.094)	(0.096)	(0.132)
Covariates	∅	∅	∅	∅
Country F.E.	∅	∅	∅	∅
Period F.E.	✓	✓	✓	✓
Period	3-Year	Year	Year	Year
Observations	606	3,125	1,994	3,277
R^2	0.074	0.042	0.188	0.011
		B. LMIC to UMIC		
Graduated	0.090	0.115	0.365**	0.235*
	(0.222)	(0.101)	(0.181)	(0.141)
Reverse graduated		−0.103	0.012	0.340
		(0.118)	(0.189)	(0.249)
Log GNIpc change	−0.185*	−0.023	1.254***	0.085
	(0.133)	(0.141)	(0.184)	(0.167)
Constant	−0.196***	0.032	−0.314***	−0.137
	(0.063)	(0.094)	(0.097)	(0.133)
Covariates	∅	∅	∅	∅
Country F.E.	∅	∅	∅	∅
Period F.E.	✓	✓	✓	✓
Period	3-Year	Year	Year	Year
Observations	606	3,125	1,994	3,277
R^2	0.073	0.042	0.190	0.012

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

Note: Dependent variables are change scores between year *t* and year *t*-1 and are standardized for ease of comparison. Independent variables (graduation indicators and change in GNI per capita) refer to the transition between year *t*-1 and year *t*-2. In the aid regressions, the sample is restricted to countries that have ever benefited from IDA after 1987. Since change scores are used for the dependent variable, no imputations are made in the creditworthiness variable.

classifications. This explanation in fact supports the strategic logic, which is rooted in the idea that donors must defend the allocation of scarce resources.

S7 Replication of Knack et al.

Knack et al. (2014) look at the effect of crossing the operational cutoff on aid disbursements from the OECD. They use data from 1987-2010 for all countries that either were IDA eligible or crossed the operational cutoff during this period and group country-years into three-year periods that correspond with the IDA replenishment cycles. When aggregating to a period, they take the income and threshold observation for the final year in the period, and for all other variables they take the mean of the three observations in the period.

There are three main differences between their model and data and mine:

1. They include Country Policy and Institutional Assessment (CPIA) index as a control. This is a score the World Bank awards to countries to evaluate economic policies, and this index is an important component in the formula that determines *allocation* of IDA, although it is not used in eligibility decisions. While the CPIA's modern equivalent the IDA Resource Allocation Index (IRAI) is available publicly after 2006, the CPIA before 2006 is private data. Nonetheless, in correspondence with the authors, they confirmed that their results are robust to dropping this control.

2. The Knack et al. data end in 2010, which is the last year of the IDA15 replenishment cycle. My data end in 2015, which allows me to create observations for IDA16 and IDA17 as well.

3. There are minor differences in the observations included in the sample, mostly due to data availability. These observations are reported in Table S9.

In Table S10, I replicate the main results of the Knack et al. paper on the data provided to me by the authors and on my data. Model 1 is the benchmark result reported in Knack et al., although without the CPIA control. Even without this control, the result is significant: crossing the operational cutoff causes total ODA to decrease. When I run the same model on my own data set in Model 2, I obtain very similar results, even though I am missing observations that appear in the first column of Table S9. In Model 3, I expand the sample to include the observations that appear in the second column of Table S9. These are the observations that meet the criteria for inclusion in the Knack sample, but were not in the Knack data set. The results continue to hold.

In Model 4, however, I include two additional replenishment cycles for which data are now available. In practice, this extends the sample from 2010 to 2016. (I continue to include my observations from Model 2.) Adding more recent data causes the result to disappear entirely.

Table S9: Differences in included observations

In Knack et al. only	In Dolan only
Afghanistan—14	Bolivia—9
Afghanistan—15	Cameroon—9
Albania—10	Congo, Rep.—9
Angola—10	Honduras—9
Bhutan—9	Maldives—9
Bhutan—11	Papua New Guinea—9
Bhutan—12	Philippines—9
Bosnia and Herzegovina—12	Samoa—9
Congo, Dem. Rep.—9	Sao Tome and Principe—9
Congo, Dem. Rep.—10	Solomon Islands—9
Congo, Dem. Rep.—12	Tanzania—9
Congo, Dem. Rep.—13	Azerbaijan—10
Djibouti—11	Cameroon—10
Djibouti—12	Congo, Rep.—10
Dominica—9	Papua New Guinea—10
Dominica—10	Samoa—10
Dominica—11	Sao Tome and Principe—10
Equatorial Guinea—13	Turkmenistan—10
Equatorial Guinea—14	Ukraine—10
Eritrea—11	Haiti—11
Gambia, The—10	Papua New Guinea—11
Gambia, The—11	Sao Tome and Principe—11
Grenada—10	Syrian Arab Republic—11
Grenada—11	Ukraine—11
Kenya—9	Haiti—12
Liberia—12	Marshall Islands—12
Liberia—13	Micronesia, Fed. Sts.—12
Mongolia—10	Sao Tome and Principe—12
Montenegro—14	Haiti—13
Nicaragua—9	Marshall Islands—13
Nicaragua—12	Micronesia, Fed. Sts.—13
Nicaragua—13	Sao Tome and Principe—13
Serbia—13	Marshall Islands—14
Serbia—14	Micronesia, Fed. Sts.—14
Somalia—9	Sao Tome and Principe—14
South Africa—11	Marshall Islands—15
South Africa—12	Micronesia, Fed. Sts.—15
South Africa—13	
South Africa—14	
South Africa—15	
St. Kitts and Nevis—9	
St. Kitts and Nevis—10	
St. Kitts and Nevis—11	
St. Lucia—10	
St. Lucia—11	
St. Vincent and the Grenadines—9	
St. Vincent and the Grenadines—10	
St. Vincent and the Grenadines—11	
Sudan—10	
Sudan—11	
Turkmenistan—11	
Turkmenistan—12	
Turkmenistan—14	
Turkmenistan—15	
Vietnam—9	
Vietnam—10	
Yemen, Rep.—10	
Zimbabwe—13	
Zimbabwe—14	

Given that Knack et al.'s results replicate perfectly but are highly sensitive to the inclusion of data published after their study, I also investigate some of their other findings, which are relevant for my argument. The authors also investigate the effects of the other thresholds that are included in my study: the LIC ceiling, the historical cutoff, and the LMIC ceiling. They do so by including these other dummies in the regression alongside the operational cutoff and find that each of these three other thresholds is not significant.

In Table S11, I replicate their analysis on their sample (my data) and on my extended sample. None of the other thresholds have any significance in the sample ending with IDA15. However, the results change remarkably when we include observations through IDA17. Although the LIC ceiling remains insignificant, the historical cutoff approaches significance, and the LMIC ceiling is highly significant. These results hold

Table S10: Replication using alternative samples

	Log ODA from all donors			
	(1)	(2)	(3)	(4)
Above operational cutoff (0-1)	−0.21**	−0.28***	−0.22**	−0.13
	(0.10)	(0.10)	(0.10)	(0.09)
Log GNI per capita (lagged)	−0.09	−0.04	−0.03	−0.17
	(0.14)	(0.09)	(0.09)	(0.11)
Log pop (lagged)	−0.19	−0.04	0.14	0.09
	(0.40)	(0.41)	(0.41)	(0.39)
Political rights (1-7, lagged)	0.04	0.05	0.04	0.05
	(0.03)	(0.03)	(0.03)	(0.04)
Constant	25.43***	6.55	3.89	7.50
	(6.98)	(6.02)	(5.99)	(6.73)
Observations	550	484	520	699
Country F.E.?	Yes	Yes	Yes	Yes
Period F.E.?	Yes	Yes	Yes	Yes
Data source	Knack et al.	Dolan	Dolan	Dolan
Extent of sample	Thru IDA15	Thru IDA15	Thru IDA15	Thru IDA17
Observations included thru IDA15	Both+Knack	Both	Both+Dolan	Both+Dolan

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

regardless of whether the operational cutoff dummy is also included.

Evidently, how donors respond to these thresholds changes over time. One possible explanation for this is that the graduation process has become increasingly flexible in its application over time. In other words, there are more countries that cross the operational cutoff but do not change lending categories. If observers respond to lending categories, then the "treatment" of crossing the operational cutoff may attenuate. An alternative explanation is that other donors are less susceptible to the bias than they once were.

Table S11: Replication using other cutoffs

	Log ODA from all donors								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Above operational cutoff (0-1)	−0.26**	−0.18		−0.29***	−0.13		−0.28***	−0.18**	
	(0.11)	(0.12)		(0.10)	(0.09)		(0.10)	(0.08)	
Above LIC ceiling (0-1)	−0.05	0.08	−0.01						
	(0.11)	(0.12)	(0.08)						
Above his cutoff (0-1)				−0.14	−0.28	−0.28			
				(0.11)	(0.18)	(0.18)			
Above LMI ceiling (0-1)							0.02	−0.42***	−0.37***
							(0.19)	(0.13)	(0.13)
Log GNI per capita (lagged)	−0.02	−0.19*	−0.23**	−0.001	−0.09	−0.14	−0.04	−0.10	−0.19**
	(0.09)	(0.11)	(0.11)	(0.09)	(0.13)	(0.11)	(0.09)	(0.09)	(0.08)
Log pop (lagged)	−0.05	0.08	0.13	−0.12	−0.09	−0.05	−0.04	−0.07	0.01
	(0.41)	(0.39)	(0.39)	(0.40)	(0.41)	(0.40)	(0.41)	(0.37)	(0.37)
Political rights (1-7, lagged)	0.05*	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)
Constant	6.57	7.78	7.09	7.57	9.97	9.61	6.49	9.67	8.89
	(6.01)	(6.79)	(6.78)	(5.94)	(6.94)	(6.92)	(6.05)	(6.28)	(6.38)
Observations	484	699	699	484	699	699	484	699	699
Country F.E.?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Period F.E.?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Extent of sample	Thru IDA15	Thru IDA17	Thru IDA17	Thru IDA15	Thru IDA17	Thru IDA17	Thru IDA15	Thru IDA17	Thru IDA17

Standard errors in parentheses
*p<0.1; **p<0.05; ***p<0.01

S8 Revisions to Income Data

I present evidence that countries manipulate their GNI per capita as they approach the thresholds that delineate categories.

S8.1 Approach

According to my theory, whether a country seeks a higher or a lower classification should depend on how much that country values aid relative to improved investment or borrowing opportunities. However, testing this theory would involve imposing assumptions about how the diverse interests of government officials and various domestic groups are aggregated into a unitary foreign policy. Since this preference aggregation process is not obvious, I leave it to future research to explain which objectives states pursue, using this space primarily to explain how they can manipulate classifications and that they do.

Some studies even show that statistical manipulation takes place at relevant thresholds. Previous works especially germane to my study have used the McCrary statistical density test to demonstrate systematic underreporting of GNI per capita in an attempt to remain below the World Bank's operational threshold.[14] These tests, however, are misleading for two reasons. First, they cannot illustrate heterogeneous strategies. I have argued that the classifications are associated with both costs and benefits, suggesting that certain countries may prefer to be under-classified, while others may prefer to be over-classified. If countries try to move in different directions, these effects could cancel each other out and present the illusion of continuity, or small discontinuities may underestimate the extent of manipulation occurring in both directions. Second, these tests will underestimate strategic behavior if some attempts fail. For example, if a country close to a threshold tries to jump just over it but fails, it can easily appear to be a country that looks like a country that tried to slide just under the threshold. Since countries are operating in a tight range, it is likely that failures are not uncommon.

S8.2 Data

Revisions to GNI are frequently significant enough to influence a country's classification. I compare the classifications that countries received based on the best available estimates at the time with the estimates that they would have received given our estimates today. In Table S12, I find 288 country-years spanning dozens

[14] Kerner et al. (2015) observe significantly more observations just below the operational threshold in the historical data from the World Bank Atlas, but not in the revised figures currently found in the WDI.

of countries are "misclassified," that is, the classifications they received would have been incorrect given the current state of knowledge. This alone is not sufficient evidence of manipulation, as such revisions could be perfectly benign, but it does point to the significance of income revisions in determining classifications.

Table S12: "Mis"-classifications

De facto	Ex post	Country-years	Countries
Underestimated economies			
L	LM	73	30
LM	UM	106	43
UM	H	49	19
		228	92
Overestimated economies			
LM	L	32	17
UM	LM	22	11
H	UM	6	1
		60	29

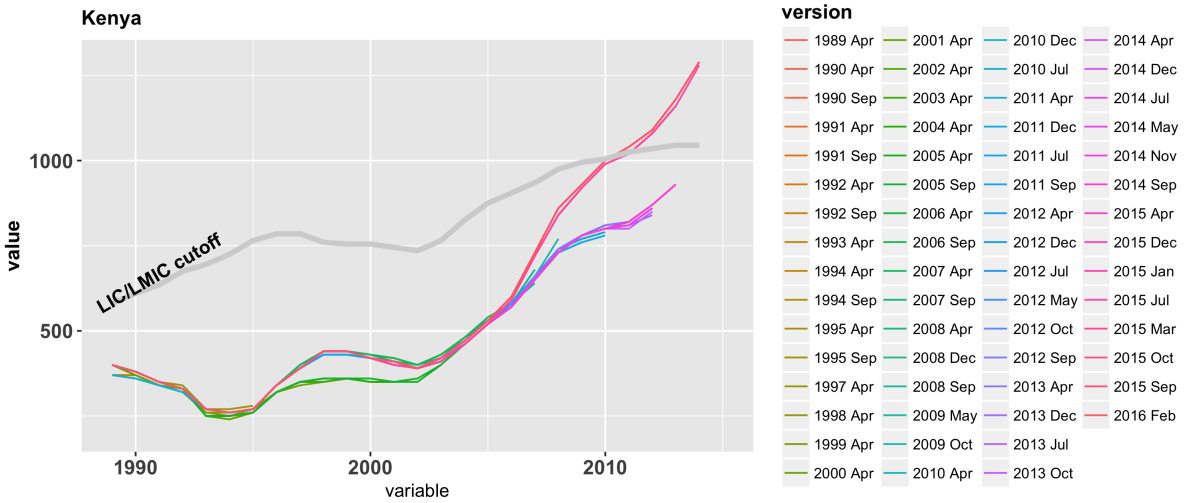
Figure S3 illustrates the unique variation reported in this "time machine" data set. Each panel plots a different country's income over time according to every version of World Bank data that has ever been published, with each version in a different color. Wherever there is vertical distance between colored lines, ex post estimates of an economy's national income have been revised in a version of the WDI indicated by the color. In these examples, the estimates diverge dramatically in the 2010s, when the countries rebased and revised their estimates of previous years ex post. The fact that different colors of lines straddle the cutoff separating LICs and LMICs indicates that these countries would have been classified differently had researchers known what we know now.

Intuitively, my objective is to calculate the magnitude of the revisions reported in each of the versions of the WDI (the area between one colored line and the colored line from the previous version of the WDI) and to relate that magnitude to the country's distance from a threshold.

S8.3 Empirical Strategy

To operationalize "revisions to GNI data," I start by calculating, for each version of the WDI, how much a country's estimates of the previous three years of GNI differed from its estimates of those same years as reported by the previous version of the WDI. Mathematically, I define a revision for each version as the

Figure S3: National income data is revised over time: the example of Kenya



Note: Figure depicts Kenya's GNI per capita over time according to every available version of the World Development Indicators, each depicted by a line with a different color. The x-axis represents the year described by the data, while the color of the line indicates the version of the data. The separation between lines occurring in the early 2010s results from a statistical rebasing exercise in which Kenya re-estimated the size of its economy. The grey line indicates the the ceiling separating LICs and LMICs, which changes over time only to account for inflation. *Source*: World Bank Database Archives, OGHIST.

following:

$$ThreeYrRevision_{i,j} = (GNI_{i,j} - GNI_{i-1,j}) + (GNI_{i,j-1} - GNI_{i-1,j-1}) + (GNI_{i,j-2} - GNI_{i-1,j-2}) \quad (2)$$

where i denotes a version of the WDI and j denotes the year described by the data. Since my "treatment," a country's distance to the threshold, is measured only annually and I am interested only in the total amount of revision carried out in a single year, I take the sum of all revisions reported in all versions in a given year. To match the data and classification schedule of the World Bank, I aggregate by fiscal year, beginning in July. Intuitively, this is a measure of the total dollars a country added or subtracted to its estimates of its GNI in the three most recent years. Since these are the years that will most likely be affected in a country's attempts to engineer its GNI for the present year, they are a good basis for a proxy for a country's revision activity in a given year. Finally, because this paper highlights reasons why countries may choose to either seek or avoid certain classifications, I remain agnostic about the direction of the revision by taking the absolute value, resulting in the following measure of revision activity:

$$ThreeYrRevisionAbs_j = |\Sigma ThreeYrRevision_{i,j}| \quad (3)$$

With this measure defined, I proceed to investigate whether a country's revision activity increases as it approaches an income classification cutoff. To do so, I make use of the analytic tools developed for regression discontinuity designs, which allow me to test whether data just before and just after a cutoff exhibit significantly different patterns. Regression discontinuities are typically employed in the service of causal identification. In these designs, an author makes use of a situation in which some "treatment" or intervention occurs only after a certain threshold is met. If this threshold is truly arbitrary, and if actors do not behave strategically, then in expectation, units that have just met and have just failed to meet the threshold should be similar. Using this design, a researcher will test whether outcomes for units just above and just below the threshold are significantly different, and if they are, this may be causally attributed to the intervention. My use of this design differs, since I am using it to illustrate that units' behavior *is* strategic.

Specifically, I estimate the size of the jump in revision activity at the discontinuity of the threshold through the use of local linear regression.[15] This approach models linear relationships on either side of the threshold and estimates the difference between them. Since regression discontinuities model the difference only in the neighborhood of the cutpoint, distant observations are omitted. Selection of the bandwidth within which observations are included, therefore, is an important modeling decision. There is some debate in the literature whether bandwidths should be selected through a data-driven process or through researcher discretion; I simply estimate the results at several reasonable bandwidths to demonstrate the sensitivity of the results.[16] In order to cluster standard errors by country, I use block bootstrapping, since there are fewer than 50 clusters within these bandwidths. Using the form

$$Y_{it} = \alpha + \beta AboveCutoff_{it} + \delta DistanceToCutoff_{it} + \gamma AboveCutoff * DistanceToCutoff_{it} + \varepsilon_{it} \quad (4)$$

I estimate β, which is equivalent to the size of the jump at the discontinuity.

S8.4 Robustness

[15] For methodological reasons to use local linear regression instead of the use of higher-order polynomials, see Skovron and Titiunik (2015).

[16] See Calonico et al. (2014).

Table S13: Discontinuities in revisions to national income data (removing outliers)

	Absolute revisions to GNIpc in previous 3 years		
	(1)	(2)	(3)
	LIC/LMIC Discontinuity		
Above cutoff	−69.89**	−39.60**	−37.86**
	(31.49)	(18.20)	(17.86)
Outliers removed?	None	Above 500	Above 300
Observations	218	216	214

Note: Estimates come from a local linear regression of the outcome on the above-cutoff indicator, the running variable (GNIpc distance to the cutoff), and the interaction. All models use a bandwidth of 100. All standard errors are calculated using block bootstrapping, clustered by country.

References

Mark. T. Buntaine, Bradley C. Parks, and Benjamin P. Buch. Aiming at the Wrong Targets: The Domestic Consequences of International Efforts to Build Institutions. *International Studies Quarterly*, 61(2):471–488, 2017.

Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326, 2014.

Allison Carnegie and Cyrus Samii. International Institutions and Political Liberalization: Evidence from the World Bank Loans Program. *British Journal of Political Science*, (Forthcoming), 2017.

Neil Fantom and Umar Serajuddin. The World Bank's Classification of Countries by Income. *World Bank Policy Research Working Paper*, 7528, 2016.

Dan Honig. When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation. *International Organization*, 2019.

Andrew Kerner, Morten Jerven, and Alison Beatty. Does It Pay to Be Poor? Testing for Systematically Underreported GNI Estimates. *The Review of International Organizations*, pages 1–38, 2015.

Stephen Knack, Colin Xu, and Ben Zou. Interactions among Donors' Aid Allocations: Evidence from an Exogenous World Bank Income Threshold. *World Bank Policy Research Working Paper*, 7039, 2014.

Helen Milner. Why Multilateralism? Foreign Aid and Domestic Principal-Agent Problems. In Darren Hawkins, David Lake, Daniel Nielsen, and Michael J. Tierney, editors, *Delegation and Agency in International Organizations*. Cambridge University Press, 2006.

Lynge Nielsen. How to Classify Countries Based on Their Level of Development: How It is Done and How It Could be Done. *IMF Working Paper*, 2011.

Christopher Skovron and Rocio Titiunik. A Practical Guide to Regression Discontinuity Designs in Political Science. 2015.

World Bank IDA Resource Mobilization Department. Review of IDA's Graduation Policy. 2016.