Supplementary Materials to B.J.Pol.S. **XX,** X–XX *Cambridge University Press, 2018*
doi:doi:10.1017/XXXX

Supplementary Materials:
Do Voters Benchmark Economic Performance?

VINCENT AREL-BUNDOCK, ANDRÉ BLAIS AND RUTH DASSONNEVILLE*

*vincent.arel-bundock@umontreal.ca, andre.blais@umontreal.ca, ruth.dassonneville@umontreal.ca

Two Equivalent Models of Benchmarking

Kayser and Peress (2012) estimate this model:

$$V = \theta_{y-i}(G_y - G_i) + \theta_i G_i + \Gamma\Omega + \varepsilon.$$

We recommend this simpler model:

$$V = \delta_y G_y + \delta_i G_i + \Gamma\Omega + \varepsilon.$$

In the main text, we explain why the second model, which excludes duplicate regressors, is preferable. In Table A1, we show numerically that the two models are logically equivalent. In particular, we see that $\theta_{y-i} \equiv \delta_y$, that $\theta_i - \theta_{y-i} \equiv \delta_i$, and that all the other coefficients are exactly identical in the two models.

TABLE A1: The simple and complicated models of benchmarking are equivalent.

	KP	ABD
Local Growth ($G_y - G_i$)	0.611774***	
	(0.234206)	
Domestic Growth (G_y)		0.611774***
		(0.234206)
Global Growth (G_i)	0.050281	-0.561493
	(0.374609)	(0.423542)
Local Unemployment	-0.040779	-0.040779
	(0.187222)	(0.187222)
Global Unemployment	-0.327835	-0.327835
	(0.216280)	(0.216280)
Coalition size	-3.332975***	-3.332975***
	(0.714253)	(0.714253)
Eff.Num.Parties	-2.774487***	-2.774487***
	(0.599097)	(0.599097)
Population	0.000039**	0.000039**
	(0.000015)	(0.000015)
Year	0.034884	0.034884
	(0.056109)	(0.056109)
Constant	-16.505970	-16.505970
	(111.444462)	(111.444462)
Observations	189	189
R^2	0.516	0.516

Standard errors in parentheses

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

Benchmarking Across Borders

In the text, we evaluated KP's evidence according to one criterion: There is evidence of benchmarking if the marginal effect of international growth, controlling for domestic growth, is negative. This condition is exactly identical to the hypothesis that KP call "partial benchmarking" in their article. In Tables A2 to A5, we present full replication results for all of KP's models. Those tables show that there is essentially no evidence of partial benchmarking in KP's own models and data.

Here, we consider three additional empirical claims from *Benchmarking Across Borders*.

Claim #1: A statistically insignificant estimate of θ_i constitutes evidence of "full benchmarking."

KP write:[1]

> If "voters fully benchmark, we would expect the coefficient on local growth to be positive and the coefficient on global growth to be zero."

In terms of Model 3, this hypothesis can be translated as $\theta_{y-i} > 0$ and $\theta_i = 0$. The $\theta_{y-i} > 0$ condition is trivial because θ_{y-i} is the marginal effect of domestic growth; both benchmarking and conventional theories make the same prediction with respect to that parameter. The full benchmarking criterion thus boils down to $\theta_i = 0$.

Unfortunately, tests of full benchmarking cannot add much credence to the theory, because they suffer from three fundamental flaws.

1. Full benchmarking is a more restrictive special case of partial benchmarking. If full benchmarking ($\theta_{y-i} > 0$ and $\theta_i = 0$), then partial benchmarking ($\theta_{y-i} > \theta_i$). The contrapositive: If no partial benchmarking, then no full benchmarking. We have shown that KP's data offer little evidence in support of the *weak* partial benchmarking condition. Thus, readers should be extremely wary of claims that the *stronger and more restrictive* full benchmarking condition finds empirical support.

2. As KP concede (footnote 11), their tests of full benchmarking are under-powered. This is because the benchmark measure of global growth was constructed by combining growth rates from many countries. A natural side-effect of this strategy is that KP's measure of global growth has low variance, and that confidence intervals around the θ_i parameter are very large. Low statistical power and high uncertainty are not problems *per se*, but they become vexing when combined with the third flaw of the full benchmarking hypothesis.

3. The $\theta_i = 0$ proposition does not offer a well-founded statistical test of benchmarking, because it would require that the analyst *accept a null hypothesis*. Accepting the null when our model fails to reject it is a well-known statistical fallacy. That trap is especially dangerous when our tests are under-powered and confidence intervals are wide.[2] If the proposed test was valid, scientists could conclude that any null hypothesis is true, provided that their sample was small enough, or that their independent variable showed sufficiently low variation. At the very least, proponents of benchmarking who hope to convince that

[1] Kayser and Peress 2012, 668.

[2] Most introductory econometrics texts discuss this issue. See for example Wooldridge (n.d.) or the amusing discussion of the consequences of "micronumerosity" (i.e., small samples) for null hypothesis tests in Goldberger (n.d.).

> $\theta_i = 0$ should demonstrate that the confidence intervals for that parameter are narrow and centered around zero. In fact, they are not.

In short, finding a statistically insignificant estimate of θ_i should not be treated as evidence of benchmarking, because that result is overdetermined by the under-powered nature of KP's test. Furthermore, since "full benchmarking" is a more restrictive special case of "partial benchmarking", any claim that there is evidence of the former but not the latter would be dubious. Thus, we must reject the "full benchmarking" hypothesis, and focus on the criterion that we used in our paper.

Claim #2: The substantive effect of "decomposed" growth is more important than the effect of domestic growth.

The foregoing discussion was largely focused on statistical significance, ignoring the information about substantive importance which can be gleaned from coefficient sizes. In their article, KP estimate a version of Equation 1, and compare the G_y coefficient in that model to the $(G_y - G_i)$ coefficient in Equation 3:

> "the difference in the effect sizes between column (1) and column (3) (our preferred benchmarking model) is worth noting. According to model (1), a 1% increase in growth is associated with a 0.604% increase in the leader party's vote share. According to model (3), a 1% increase in [the local component of] growth is associated with a 1.261% increase in the leader party's vote share. The estimated effect size therefore more than doubles when moving from the conventional model to the benchmarked model."

As we explained in the text, the coefficient associated with the $G_y - G_i$ composite variable is the marginal effect of domestic growth on votes for the incumbent; it has no meaningful interpretation in terms of the effect of relative economic performance on votes. As such, a comparison between β_y and θ_{y-i} seems to be of little relevance.

Moreover, we note that the coefficient estimate described in the quote above is extreme. In the 12 aggregate-data models that KP consider, the coefficient estimate associated with $G_y - G_i$ takes on values of: 1.261, 0.612, 0.529, 0.646, 0.818, 0.820, 0.538, 0.683, 0.474, 0.687, 0.582, 0.756. The first value is clearly an outlier which, incidentally, was produced by the only model (out of 24) that allowed KP to reject the null hypothesis of "no benchmarking."

Claim #3: At several points in time, the magnitude of the benchmarked economic vote is greater than the magnitude of the non-benchmarked economic vote.

For Figure 2 of their article, KP estimate Models 1 and 2 over different time windows, and plot the evolution of the β_g and λ_{g-r} coefficients over the 1990-2010 period. Considering that KP explicitly reject Model 2 as an inappropriate test of benchmarking,[3] this comparison seems irrelevant. Moreover, the models used in Figure 2 lack the control variables that make KP's core model specification more credible.

[3] Kayser and Peress 2012, 663.

Relative Economic Performance and the Incumbent Vote

Aytaç's additive specification

In our article, we estimated a series of models closely related, but not exactly identical, to those used by Aytaç.[4] In his article, Aytaç estimates this model (Table 1, Model 2):

$$V = \kappa_{th}(G_t - G_h) + \kappa_{ti}(G_t - G_i) + \kappa_i G_i + \kappa_y G_y + \Lambda\Omega + \nu \qquad (7)$$

The author defines "Relative Domestic Growth" as $G_t - G_h$ and "Relative International Growth" as $G_t - G_i$. Then, he interprets the results as follows:

> Model 7 "introduces my measures of relative performance, Relative Domestic Growth and Relative International Growth. Both variables have positive and statistically significant effects on the incumbent's vote, providing evidence for the hypothesis that voters reward (punish) incumbents on whose watch the economy performs relatively better (worse) in domestic and international comparisons."[5]

This is precisely the mistake that we warned against in our article. When we control for "Relative International Growth" ($G_t - G_i$), the coefficient associated with "Relative Domestic Growth" ($G_t - G_h$) loses its obvious substantive interpretation. κ_{th} does not measure benchmarking. Likewise, by "partialling-out" one of the bounds of the $G_t - G_i$ interval, we change the substantive meaning of its associated coefficient. As a result, κ_{ti} no longer has a straightforward interpretation in terms of the gap between domestic and international growth.

As we explained in our article, the correct way to test benchmarking is to look at the marginal effects of our variables of interest. If voters engage in benchmarking, we should find:

$$\frac{\partial V}{\partial G_t} = \kappa_{th} + \kappa_{ti} > 0 \qquad \text{(Marginal effect of full term growth)} \qquad (8)$$

$$\frac{\partial V}{\partial G_y} = \kappa_y > 0 \qquad \text{(Marginal effect of election year growth)} \qquad (9)$$

$$\frac{\partial V}{\partial G_i} = \kappa_i - \kappa_{ti} < 0 \qquad \text{(Marginal effect of international growth)} \qquad (10)$$

$$\frac{\partial V}{\partial G_h} = -\kappa_{th} < 0 \qquad \text{(Marginal effect of historical growth)} \qquad (11)$$

Based on these criteria, the results in the original article no longer reject the null hypothesis of "no international benchmarking", but there is still support historical benchmarking. This is consistent with the results we reported in our article, where we concluded that none of the models showed evidence of international benchmarking, but that there seemed to be evidence of historical benchmarking in the models where voters are assumed to rationally consider the average level of growth over the incumbent's full term in office.

The only minor difference between the results that Aytaç reports[6] and ours, is due to the fact that he includes both G_y and G_t in the same model, whereas we treat those two measures of

[4] Aytaç 2018.
[5] 24.
[6] Aytaç 2018.

domestic growth as competing alternatives, in separate models. We believe that our approach is more principled: Given that the election year is part of the incumbent's term in office, it makes little sense to talk about the effect of changes in G_y, holding G_t constant. Still, we concede that reasonable people could disagree on this point.

Aytaç's multiplicative specification

To test conditional benchmarking, Aytaç estimates a model of this form:

$$\begin{aligned} V =& \kappa_{ti}(G_t - G_i) + \kappa_{mti} M(G_t - G_i) + \\ & \kappa_{th}(G_t - G_h) + \kappa_{mth} M(G_t - G_h) + \\ & \kappa_y G_y + \kappa_{my} M G_y + \\ & \kappa_i G_i + \kappa_m M + \Lambda\Omega + \nu, \end{aligned} \tag{12}$$

where M stands for a variable that moderates comparative economic assessments.

By now, the next step should be familiar. If voters benchmark, these conditions should hold:

$$\frac{\partial V}{\partial G_t} = \kappa_{ti} + \kappa_{th} + (\kappa_{mti} + \kappa_{mth})M > 0 \qquad \text{(Marginal effect of full term growth)} \tag{13}$$

$$\frac{\partial V}{\partial G_y} = \kappa_y + \kappa_{my} M > 0 \qquad \text{(Marginal effect of election year growth)} \tag{14}$$

$$\frac{\partial V}{\partial G_i} = \kappa_i - \kappa_{ti} - \kappa_{mti} M < 0 \qquad \text{(Marginal effect of international growth)} \tag{15}$$

$$\frac{\partial V}{\partial G_h} = -\kappa_{th} < 0 \qquad \text{(Marginal effect of historical growth)} \tag{16}$$

The equations above highlight some of the pitfalls of using composite variables to test conditional theories of benchmarking. Indeed, the marginal effects in Equation 12 are non-trivial expressions. In the original article, Aytaç (2018) fails to consider this complexity: He plots $\kappa_{ti} + \kappa_{mti} M$ and assesses the evidence of benchmarking on the basis of that graph. However, Equation 15 makes clear that this is not actually the quantity of interest. Specifically, the expression that Aytaç considers is missing the κ_i component. And since all the estimates of κ_i in his model are positive, Aytaç's plot systematically overstates the evidence of international benchmarking.

TABLE A2: Replication of KP Table 1 – Aggregate-level results for benchmarking in the economic vote

	Model 1	Model 2 Median	Model 3 PC	Model 4 Trade
$G_{d\,t-1}$	0.6044	0.8185	1.2612	0.8196
	(0.2673)	(0.3315)	(0.3525)	(0.4037)
$G_{g\,t-1}$		-0.7207	-1.3056	0.0846
		(0.6019)	(0.4661	(0.7039)
$\text{Unem}_{d\,t-1}$	-0.2485	-0.2095	-0.3359	0.0274
	(0.2055)	(0.2095)	(0.2287)	(0.2113)
$\text{Unem}_{g\,t-1}$		-0.3806	0.2058	-0.3046
		(0.5268)	(0.2554)	(0.5433)
(Intercept)	34.0951	37.4457	35.0705	32.7637
	(1.6989)	(3.9438)	(1.9972)	(4.3722)
N elections	213	213	213	146
N countries	22	22	22	22
R^2	0.029	0.035	0.061	0.034
Adj. R^2	0.020	0.017	0.043	0.007
RMSE	10.4858	10.5025	10.3616	10.4145

OLS estimates with heteroskedasticity robust standard errors in parentheses.

TABLE A3: Replication of KP Table 3 – Robustness checks for aggregate-level models

	Model 1 Median	Model 2 PC	Model 3 Trade	Model 4 Median	Model 5 PC	Model 6 Trade	Model 7 Median	Model 8 PC	Model 9 Trade
$G_{d\,t-1}$	0.5375	0.6118	0.6826	0.4744	0.5288	0.6869	0.5824	0.6360	0.7561
	(0.2184)	(0.2342)	(0.2844)	(0.1902)	(0.2163)	(0.2304)	(0.2264)	(0.3235)	(0.2970)
$G_{g\,t-1}$	-0.4402	-0.5615	-0.2326	-0.4690	-0.3837	-0.3442	-0.3578	-0.2741	-0.0250
	(0.6113)	(0.4235)	(0.6307)	(0.4794)	(0.3819)	(0.4850)	(0.4682)	(0.4815)	(0.4581)
$Unem_{d\,t-1}$	-0.1480	-0.0411	-0.0653	-0.2413	-0.2524	-0.1812	0.0025	0.1848	-0.1581
	(0.1610)	(0.1872)	(0.1584)	(0.1333)	(0.1686)	(0.1388)	(0.2030)	(0.2782)	(0.2734)
$Unem_{g\,t-1}$	-0.3196	-0.2868	-0.1552	-0.4388	-0.0748	-0.1757	-0.6272	-0.6649	-0.3462
	(0.4228)	(0.2265)	(0.4309)	(0.3323)	(0.2104)	(0.4065)	(0.4139)	(0.4539)	(0.6718)
coalsize	-3.2305	-3.3331	-3.6446	-1.3574	-1.5058	-1.7489	-1.4885	-1.3986	-1.5110
	(0.7206)	(0.7143)	(0.9056)	(0.6003)	(0.5941)	(0.7101)	(0.5393)	(0.6114)	(0.6041)
enep	-2.7303	-2.7742	-2.0543	0.4177	0.4537	0.5802	2.3533	2.4167	3.4912
	(0.6086)	(0.5991)	(0.6761)	(0.5620)	(0.5481)	(0.6896)	(0.8824)	(0.8758)	(1.0286)
pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0002	0.0006
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0006)
elecyr	0.0559	0.0349	0.0761	0.0258	0.0056	-0.0293	-0.0526	-0.0455	-0.1427
	(0.0591)	(0.0561)	(0.1033)	(0.0467)	(0.0428)	(0.0664)	(0.0577)	(0.0625)	(0.1108)
Vote (previous election)				0.7753	0.7650	0.7284	0.7697	0.7562	0.8054
				(0.0770)	(0.0765)	(0.0959)	(0.0739)	(0.0842)	(0.1163)
(Intercept)	-58.1209	-16.5549	-103.3335	-41.2479	-2.8184	67.4097	99.8943	85.4243	263.3620
	(116.7971)	(111.4470)	(206.5107)	(91.8039)	(84.4547)	(131.1914)	(110.7509)	(119.6557)	(211.0814)
N elections	189	189	131	189	189	131	189	189	131
N countries	21	21	21	21	21	21	21	21	21
R^2	0.508	0.516	0.481	0.710	0.709	0.671	0.529	0.532	0.543
Adj. R^2	0.487	0.494	0.447	0.696	0.695	0.646	0.505	0.509	0.509
RMSE	7.8379	7.7784	7.8773	6.0339	6.0431	6.2998	5.5732	5.5552	5.5645

OLS estimates with heteroskedasticity robust standard errors in parentheses.

TABLE A4: Replication of KP Table 5 – Individual-level results for benchmarking in the economic vote

	Model 1	Model 2 Median	Model 3 PC	Model 4 Trade
Outside option	-1.9415	-1.9435	-1.9411	-1.9171
	(0.1939)	(0.1932)	(0.1943)	(0.1988)
Leader party	0.6706	0.9432	0.7031	1.2971
	(0.2271)	(1.6838)	(0.2454)	(0.4029)
Policy distance	-0.4984	-0.4986	-0.4984	-0.5022
	(0.0243)	(0.0236)	(0.0242)	(0.0250)
$G_{d\,t-1}$ x Leader party	0.1009	0.1139	0.1206	0.1110
	(0.0406)	(0.0396)	(0.0485)	(0.0452)
$Unem_{d\,t-1}$ x Leader party	-0.0083	-0.0036	-0.0030	0.0058
	(0.0289)	(0.0275)	(0.0296)	(0.0293)
$G_{g\,t-1}$ x Leader party		-0.1058	-0.0564	-0.0732
		(0.1087)	(0.0449)	(0.0614)
$Unem_{g\,t-1}$ x Leader party		0.0030	0.0043	-0.0794
		(0.2640)	(0.0153)	(0.0432)
N voters	40260	40260	40260	40260
N elections	31	31	31	31
N countries	17	17	17	17
pseudo-R^2	0.166	0.166	0.166	0.168

Coefficients of conditional logit models explaining vote for the lead party.
Standard errors in parentheses are clustered by election.

TABLE A5: Replication of KP Table 6 – Robustness checks for individual-level models

	Model 1 Median	Model 2 PC	Model 3 Trade	Model 4 Median	Model 5 PC	Model 6 Trade	Model 7 Median	Model 8 PC	Model 9 Trade
Leader party	0.4349	0.1991	0.5450	0.7500	0.4101	0.8585	0.6655	0.5339	0.9419
	(1.4502)	(0.1980)	(0.3953)	(1.4234)	(0.2905)	(0.4196)	(1.4788)	(0.3176)	(0.4546)
Policy distance	-0.5035	-0.5036	-0.5036	-0.5042	-0.5041	-0.5045	-0.5026	-0.5028	-0.5028
	(0.0244)	(0.0245)	(0.0261)	(0.0247)	(0.0248)	(0.0265)	(0.0245)	(0.0247)	(0.0264)
$G_{d\,t-1}$ x Leader party	0.0835	0.1305	0.1184	0.0870	0.1268	0.1172	0.0898	0.1273	0.1190
	(0.0592)	(0.0588)	(0.0530)	(0.0567)	(0.0589)	(0.0501)	(0.0586)	(0.0626)	(0.0527)
$G_{g\,t-1}$ x Leader party	-0.0425	-0.0501	-0.1006	-0.0394	-0.0510	-0.0914	-0.0268	-0.0466	-0.0803
	(0.1066)	(0.0588)	(0.0826)	(0.1028)	(0.0592)	(0.0787)	(0.1045)	(0.0607)	(0.0809)
$Unem_{d\,t-1}$ x Leader party	-0.0163	-0.0002	-0.0133	-0.0212	-0.0068	-0.0184	-0.0259	-0.0121	-0.0226
	(0.0242)	(0.0241)	(0.0201)	(0.0233)	(0.0242)	(0.0193)	(0.0231)	(0.0235)	(0.0188)
$Unem_{g\,t-1}$ x Leader party	-0.0035	-0.0228	-0.0140	-0.0166	-0.0172	-0.0215	0.0067	-0.0159	-0.0194
	(0.2089)	(0.0210)	(0.0490)	(0.2047)	(0.0217)	(0.0482)	(0.2147)	(0.0231)	(0.0492)
Additional controls									
Party characteristics	X	X	X	X	X	X	X	X	X
Demographics x Leader party				X	X	X	X	X	X
Demographics x Party characteristics							X	X	X
N voters	38177	38177	36229	37602	37602	35658	37602	37602	35658
N elections	31	31	30	31	31	30	31	31	30
N countries	17	17	17	17	17	17	17	17	17
pseudo-R^2	0.214	0.215	0.210	0.215	0.215	0.211	0.222	0.223	0.219

Coefficients of conditional logit models explaining vote for the lead party. Standard errors in parentheses are clustered by election.

TABLE A6: Conditional benchmarking. OLS regressions with country-clustered standard errors.

	Education	Education	Trade	Trade	Income	Income
G_i	0.376	0.515	0.621	0.806	-0.219	0.094
	(1.158)	(1.071)	(1.402)	(1.290)	(1.043)	(1.026)
G_i X Moderator	-1.356	-0.913	-1.578	-1.211	-0.210	-0.064
	(1.714)	(1.573)	(1.966)	(1.769)	(1.366)	(1.361)
G_t	0.864		1.350*		2.076***	
	(0.575)		(0.795)		(0.722)	
G_t X Moderator	0.811		-0.016		-1.099	
	(1.013)		(1.331)		(0.912)	
G_y		0.737*		0.762**		1.339***
		(0.373)		(0.351)		(0.417)
G_y X Moderator		-0.090		-0.107		-0.978
		(0.719)		(0.629)		(0.594)
Moderator	3.576	4.117	8.652	6.691	7.519	5.474
	(5.379)	(5.409)	(6.465)	(6.067)	(5.395)	(4.830)
Vote share lag	0.672***	0.691***	0.683***	0.698***	0.678***	0.695***
	(0.065)	(0.065)	(0.064)	(0.064)	(0.065)	(0.064)
Coalition	0.022	0.153	-0.069	0.058	0.021	0.118
	(1.112)	(1.056)	(1.058)	(1.031)	(1.105)	(1.072)
ENP	-1.557***	-1.464***	-1.600***	-1.489***	-1.546***	-1.434***
	(0.396)	(0.367)	(0.374)	(0.351)	(0.379)	(0.354)
Presidential	-4.417***	-4.272***	-4.278***	-4.237***	-3.931***	-3.987***
	(1.150)	(1.146)	(1.220)	(1.210)	(1.184)	(1.140)
Re-run	12.571***	11.841***	12.592***	11.942***	12.644***	11.773***
	(2.586)	(2.616)	(2.477)	(2.519)	(2.479)	(2.479)
Constant	9.838*	9.293*	5.794	6.988	5.974	7.308
	(5.585)	(5.301)	(5.994)	(5.698)	(5.606)	(5.084)
Observations	457	457	458	458	457	457
R^2	0.602	0.585	0.603	0.586	0.600	0.586

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

References

Aytaç, Selim E. 2018. "Relative Economic Performance and the Incumbent Vote: A Reference Point Theory." *Journal of Politics* 80 (1): 16–29.

Goldberger, Arthur Stanley. n.d. *A Course in Econometrics.* Cambrdidge (MA): Harvard University Press.

Kayser, Mark Andreas, and Michael Peress. 2012. "Benchmarking across Borders: Electoral Accountability and the Necessity of Comparison." *American Political Science Review* 106 (3): 661–684.

Wooldridge, Jeffrey M. n.d. *Introductory Econometrics: A Modern Approach.* Scarborough (ON): Nelson Education.