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

Do Voters Benchmark Economic Performance?

VINCENT AREL-BUNDOCK, ANDRÉ BLAIS AND RUTH DASSONNEVILLE*

Keywords: economic voting, benchmarking, relative evaluations, reference point, performance voting

The conventional theory of economic voting is that voters reward or punish the incumbent government based on how the domestic economy is doing. Recently, scholars have challenged that view, arguing that voters use relative assessments to gauge government performance. From this perspective, what matters is not how well the national economy is doing *per se*, but rather how it performs relative to an international or historical reference point.

This article revisits prominent published works in that emerging tradition, and finds that the available evidence does not support the benchmarking hypothesis. We come to this conclusion after taking a close look at the regression models that are typically used to test benchmarking. We show algebraically that the way in which those models are specified invites a fundamental misreading of the evidence. Finally, we propose an alternative regression equation which can be used to test benchmarking, avoids common misinterpretations, and allows us to assess complex, conditional theories of relative evaluation.

Background

Economic voting is one of the most important accountability mechanisms at work in electoral democracies. The fact that voters reward or punish the incumbent government based on how the domestic economy is doing[1] is traditionally viewed as normatively desirable, for it reflects popular control of representatives.[2]

Recently, a number of scholars have challenged this optimistic view, by pointing out that domestic economic growth is often a weak proxy for government performance. When the local economy moves in sync with secular trends or global shocks, governments may be rewarded or punished for events beyond their control.[3] This is especially true in integrated economies, where domestic fortunes are tightly linked to events abroad, and responsibility is blurred.[4] Democratic accountability thus requires more from voters than a simple response to local economic conditions.[5]

*All three authors are from the Département de Science Politique, Université de Montréal. Contact: vincent.arel-bundock@umontreal.ca, andre.blais@umontreal.ca, ruth.dassonneville@umontreal.ca. We thank Erdem Aytaç, Timothy Hellwig, Michael Lewis-Beck, seminar participants at Texas A&M University, and our regular lunch companions Chez Valère. The data, replication instructions, and the data's codebook can be found at http://dx.doi.org/10.7910/DVN/OCNIWD

[1] Lewis-Beck 1988.

[2] Fiorina 1981; Key 1966; Przeworski, Stokes, and Manin 1999.

[3] Bartels 2012.

[4] Duch and Stevenson 2010; Fernàndez-Albertos 2006; Hellwig and Samuels 2014.

[5] Achen and Bartels 2016; Anderson 2007.

A new and influential strand of research argues that voters do, in fact, make rational judgements about government performance, because their evaluations are relative.[6] What matters to rational voters may not be how well the national economy is doing *per se*, but rather how it performs relative to the economies of other countries, or relative to some historical benchmark.

Benchmarking is a powerful idea, which can be traced back to the work of Powell and Whitten. As these authors point out, voters are likely to "evaluate government relative to some expectations about how the economy should have performed".[7] But since expectations are difficult to measure, "it seems reasonable to use the international average levels of growth, inflation, and unemployment to estimate a baseline against which each country's citizens could judge the performance of their own economy." This approach is intuitive, since "abundant research in other domains of social science supports the proposition that individuals are sensitive to comparative assessments."[8]

Yet, there are also good reasons to doubt that voters benchmark economic performance. First, the benchmarking hypothesis is at odds with a dominant view on the cognitive limitations of ordinary voters. Indeed, a long tradition of research in political science has depicted the citizenry as poorly informed[9] and biased.[10] It is difficult to imagine how such an unsophisticated electorate could systematically and accurately compare how well the national economy is performing relative to other countries or past history. Second, even if some authors posit that the media could facilitate benchmarking by making implicit comparisons in their news coverage, the evidence for the underlying mechanism is rather weak. For instance, Kayser and Peress (henceforth, KP) report that high-information voters – those most exposed to media – do not engage in more benchmarking than low-information voters.[11] Finally, some empirical studies claim that voters act based on relative economic conditions, but others find that when it comes to evaluating government performance, "the effect of luck is larger than the effect of competence."[12] In short, the theoretical case for benchmarking is muddled, and the empirical record is mixed.

In this article, we show that the empirical evidence of benchmarking is extremely weak. We argue that that the way in which regression models are typically specified to test benchmarking is needlessly complicated, that it invites a fundamental misreading of the evidence, and that it often leads researchers astray. We propose a simpler model specification which can be used to test benchmarking, avoids common misconceptions, and carries powerful intuitions about the theory. We also show how this simple model can be enhanced to test more complicated theories, such as when voters benchmark against multiple reference points, or when the strength of benchmarking depends on the context. We revisit a prominent empirical study of *Benchmarking Across Borders*,[13] and conduct a faithful replication of all the models reported in that article. When correctly interpreted, the results do *not* support the contention that voters make rational comparative evaluations.

Our findings have important implications for the field of economic voting, and for our understanding of the mechanisms that underpin democratic accountability. More generally, our article makes useful contributions to political science methodology, by highlighting the shortcomings of a widely used empirical strategy, and by proposing a better way to test theories of relative evaluation.

[6] Aytaç 2018; Ebeid and Rodden 2006; Kayser and Peress 2012; Wolfers 2002.

[7] Powell and Whitten 1993, 396.

[8] Kayser and Peress 2012, 664.

[9] Converse 2000; Zaller 1992.

[10] Achen and Bartels 2016; Taber and Lodge 2006.

[11] Kayser and Peress 2012.

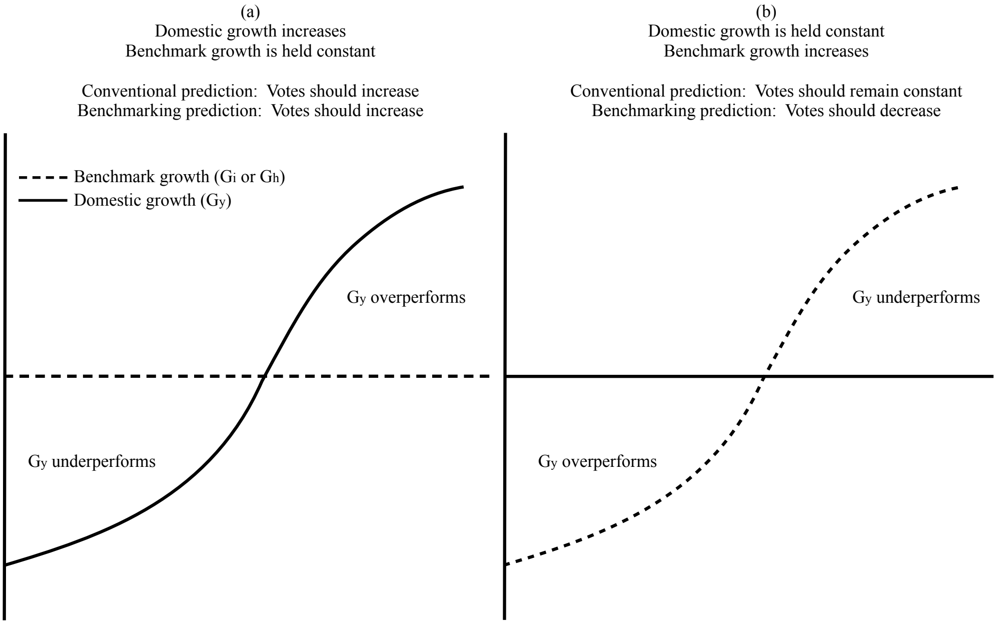
[12] Leigh 2009.

[13] Kayser and Peress 2012.

Benchmarking vs. Conventional Economic Voting

The core intuition of benchmarking is illustrated in Figure 1, where we show how domestic growth and a reference point can affect support for the incumbent party. The solid line represents the domestic growth rate during the election year (G_y), and the dashed line represents the growth rate that voters use as a benchmark to evaluate the incumbent government's performance. Depending on the analyst's theory, the reference point could be the international growth rate (G_i) or the historical level of growth in the country under study (G_h).

Figure 1: Marginal effects of domestic economic growth and benchmark growth on votes for the incumbent.


(a)
Domestic growth increases
Benchmark growth is held constant
Conventional prediction: Votes should increase
Benchmarking prediction: Votes should increase
Benchmark growth (Gi or Gh)
Domestic growth (Gy)
Gy overperforms
Gy underperforms
(b)
Domestic growth is held constant
Benchmark growth increases
Conventional prediction: Votes should remain constant
Benchmarking prediction: Votes should decrease
Gy underperforms
Gy overperforms

The conventional view of economic voting is that votes for the incumbent (V) are tied to domestic growth. As we move from left to right, G_y increases in Figure 1a but stays constant in 1b. Thus, the conventional prediction is that votes for the incumbent will increase in 1a but stay constant in 1b.[14]

In contrast, proponents of benchmarking argue that what matters to voters is not domestic growth *per se*, but rather the difference between domestic growth and the benchmark ($G_y - G_i$).[15] When the solid line is above the dashed line, domestic growth outperforms the benchmark, and voters should reward the incumbent. When the solid line is below the dashed line, domestic growth underperforms relative to the benchmark, and voters should punish the incumbent. As we move from left to right in Figure 1, the "performance gap" or "competence signal" increases in Figure 1a and decreases in 1b. Thus, benchmarking predicts that votes for the incumbent will increase in 1a and decrease in 1b.

[14] Many proponents of the conventional view would remain agnostic with respect to Figure 1b.

[15] For simplicity, we discuss international benchmarking, but if the analyst is interested in historical benchmarking, the relevant performance gap would be $G_y - G_h$.

These expectations can be restated using the language of multiple regression. When domestic growth increases and the reference point is held constant (Figure 1a), both theories predict that the incumbent's vote share will increase. In other words, both theories predict that the marginal effect of domestic growth will be positive: $\partial V / \partial G_y > 0$. When the reference point increases and domestic growth is held constant (Figure 1b), benchmarking predicts that votes for the incumbent will decrease. In other words, benchmarking predicts that the marginal effect of the reference point will be negative: $\partial V / \partial G_i < 0$.

If we hope to discriminate between benchmarking and conventional economic voting, *the main quantity of interest is the marginal effect of the reference point*, since this is where benchmarking theory makes a distinctive prediction.

How Do Scholars Test Benchmarking?

Conventional theories of economic voting are typically tested using models of this form:

$$V = \beta_y G_y + \Psi\Omega + \nu, \tag{1}$$

where V is the incumbent's vote share; G_y is the domestic economic growth rate during the election year; Ω is a vector of control variables; and ν is a disturbance term. Clearly, Model 1 cannot be used to test benchmarking, since it ignores relative evaluations altogether.

In their seminal article, Powell and Whitten[16] estimate a regression equation of this form:

$$V = \lambda_{y-i}(G_y - G_i) + \Phi\Omega + \upsilon, \tag{2}$$

with G_i equal to a "reference point", the international economic growth rate. Model 2 takes us very close to the benchmark story: When the gap between G_y and G_i is positive, the domestic economy outperforms the global economy, and voters should reward the incumbent government for its competence.

As KP note, however, Model 2 cannot be used to distinguish between benchmarking and conventional economic voting, because it suffers from omitted variable bias.[17] Indeed, the composite variable $(G_y - G_i)$ is highly correlated with the level of domestic economic growth (G_y).[18] As a result, we cannot parse out the effect of benchmarking from conventional economic voting, and λ_{y-i} captures both phenomena. Model 2 is thus useful if we want to estimate something akin to the "total effect" of domestic growth and benchmarking on voting behavior, but not if we wish to compare and contrast the two theories.

To solve this problem, KP introduce an additional control for the reference point:

$$V = \theta_{y-i}(G_y - G_i) + \theta_i G_i + \Gamma\Omega + \varepsilon. \tag{3}$$

In this model, $G_y - G_i$ represents a "decomposed" or "local" component of growth, whereas G_i aims to control for changes in the reference point. Model 3 has had tremendous influence in the field. At the time of writing, KP's article has been cited over 150 times, and several other researchers have adopted and adapted their empirical strategy.

[16] Powell and Whitten 1993.

[17] Kayser and Peress 2012, 663.

[18] In KP's dataset the correlation coefficient between G_y and $(G_y - G_i)$ equals .83.

A Widespread Misconception

An intuitive – but ultimately incorrect – way to interpret Model 3 would be to focus on the gap between G_y and G_i, and to treat the θ_{y-i} coefficient as the effect of relative economic performance on votes for the incumbent.

For example, Aytaç argues that a positive estimate of θ_{y-i} provides "evidence for the hypothesis that voters reward (punish) incumbents on whose watch the economic performs relatively better (worse) in domestic and international comparisons."[19] KP define "local growth" as the gap between G_y and G_i, and interpret θ_{y-i} as measuring the association between "an increase in local growth" and an "increase in the leader party's vote share".[20] Goplerud and Schleiter follow in those footsteps, and discuss θ_{y-i} as the effect of some "benchmarked" or "local" component of growth on voting behavior.[21] Using data on the American states, Ebeib and Rodden also interpret θ_{y-i} as the effect of "relative state conditions" on votes.[22]

If the domestic economy outperforms a reference point, it may be reasonable for voters to infer that the government is doing good work. In that spirit, Leigh treats θ_{y-i} as the effect of "government competence" on votes for the incumbent.[23,24] In the American context, Wolfers considers the gap between state and national-level economic growth, and calls θ_{y-i} the "effect of competence."[25]

Interpreting θ_{y-i} as the effect of relative economic performance on votes for the incumbent appeals to common sense, but it is a mistake. The root of the problem lies in the fact that G_i appears twice on the right-hand side of Equation 3. This redundancy changes the substantive meaning of our regression coefficients.

To see how, take the partial derivative of Equation 3 with respect to G_y, and find the marginal effect of domestic growth:

$$\frac{\partial V}{\partial G_y} = \theta_{y-i}. \tag{4}$$

This simple exercise demonstrates that the coefficient associated with $G_y - G_i$ is exactly equivalent to the marginal effect of G_y. Against intuitive common sense, θ_{y-i} does *not* measure the effect of relative economic performance on votes for the incumbent. Since θ_{y-i} is the marginal effect of domestic growth, finding a positive coefficient for "benchmarked" or "local" growth is actually supportive of conventional economic voting.

Tests of benchmarking based on Equation 3 have been repeatedly misinterpreted in prestigious scientific journals, by leading scholars of economic voting. The inclusion of duplicate regressors on the right-hand side of Equation 3 has been a source of widespread confusion in the economic voting literature.[26]

To put this confusion to rest, we need a simpler, more direct test of benchmarking.

[19] Aytaç 2018.

[20] Kayser and Peress 2012, 669.

[21] Goplerud and Schleiter 2016, 444.

[22] Ebeid and Rodden 2006, 537-539.

[23] Leigh 2009.

[24] The quantities are interpreted slightly differently in Leigh's model, since the author uses a conditional logit model, but the methodological problem remains the same.

[25] Wolfers 2002.

[26] For a short review of economic benchmarking, see Healy and Malhotra (2013), 296-297.

A Simpler Test of Benchmarking

From Figure 1, we learned that both benchmarking and conventional economic voting predict that the marginal effect of domestic growth should be positive. In contrast, only benchmarking predicts that the marginal effect of international growth should be negative. The simplest and most direct way to test those predictions is to estimate a model of this form:

$$V = \delta_y G_y + \delta_i G_i + \Gamma\Omega + \varepsilon. \tag{5}$$

Since Model 3 includes redundant regressors, it carries no more information than the simpler Model 5. In fact, Models 3 and 5 are perfectly equivalent from a logical standpoint, and they produce identical numerical results: The marginal effect of domestic growth, the marginal effect of international growth,[27] the intercept, the control variables' coefficients, the residuals, and all fit statistics are always the same in both models. In the online appendix, we present side-by-side estimates using Models 3 and 5 to illustrate this point numerically.

Yet, even if the two models are formally equivalent, the simpler specification has major advantages in terms of transparency, presentation, and interpretation.

First, the correct interpretation of KP's Model 3 is highly counter-intuitive: The coefficient that they call "Local Component of Growth" in their regression tables (θ_{y-i}) does *not* measure the effect of the local economy's relative performance on votes for the incumbent. As we showed above, this has been a major source of confusion in the field, and benchmarking results have been repeatedly misinterpreted in print. Our simpler specification avoids this problem.[28]

Second, Equation 5 directly translates the theoretical intuitions conveyed by Figure 1, and it immediately reveals the relevant test statistics. Recall that the discriminating test of benchmarking is that international growth should have a negative marginal effect on votes for the incumbent. In Model 5, the marginal effect of international growth is the δ_i coefficient, and we can simply look at its p-value. In Model 3, the marginal effect of international growth is a linear combination of coefficients ($\partial V/\partial G_i = \theta_i - \theta_{y-i}$), and we must conduct an extra Wald test to know if that combination is negative and statistically significant.

Finally, as we show below, our simpler specification offers a solid foundation on which we can build empirical tests for theories of benchmarking where voters compare multiple reference points, or where the strength of benchmarking is context-dependent.

Replication: Benchmarking Across Borders

As we explained above, the key quantity of interest for tests of benchmarking is the marginal effect of international growth (holding domestic growth constant). Unfortunately, KP do not consistently report the statistics that are needed to test if that quantity is distinguishable from zero.[29] As a result, readers cannot assess the strength of the evidence simply based on the findings printed in *Benchmarking Across Borders*.

[27] It is easy to show algebraically that $\delta_y \equiv \theta_{y-i}$, and that $\delta_i \equiv \theta_i - \theta_{y-i}$.

[28] Although KP's model includes a single variable to represent the difference between G_y and G_i, it does not offer a single parameter to measure the association between that gap and votes for the incumbent.

[29] KP only report the results of the relevant Wald test of coefficient equality for 3 of their 24 empirical models.

To see if KP's data support their theory, we re-estimated all of their models using the authors' replication files, and we computed all the quantities of interest.[30] Table 1 shows the results for four models,[31] estimated using KP's preferred measure of international growth (an index constructed via principal components analysis).

TABLE 1: OLS regression models with incumbent vote share as dependent variable.

	Baseline	Controls	Lag	Lag + FE
Domestic Growth (G_y)	1.261***	0.612***	0.529**	0.636*
	(0.352)	(0.234)	(0.216)	(0.323)
Global Growth (G_i)	-1.306***	-0.561	-0.384	-0.274
	(0.466)	(0.424)	(0.382)	(0.481)
Local Unemployment	-0.335	-0.041	-0.252	0.186
	(0.229)	(0.187)	(0.169)	(0.278)
Global Unemployment	-0.130	-0.328	-0.327*	-0.480
	(0.264)	(0.216)	(0.178)	(0.320)
Coalition size		-3.333***	-1.506**	-1.398**
		(0.714)	(0.594)	(0.611)
Eff.Num.Parties		-2.774***	0.453	2.417**
		(0.599)	(0.548)	(0.875)
Population		0.000**	0.000	0.000
		(0.000)	(0.000)	(0.000)
Year		0.035	0.006	-0.045
		(0.056)	(0.043)	(0.062)
Leader vote lag			0.765***	0.756***
			(0.077)	(0.084)
Constant	35.069***	-16.506	-2.785	85.467
	(1.997)	(111.444)	(84.454)	(119.667)
Observations	213	189	189	189

Robust standard errors in parentheses

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

Baseline specification

In column 1 of Table 1, we see that the G_y coefficient is positive. This is consistent with both benchmarking and conventional economic voting. The G_i coefficient is negative and statistically significant. This is consistent with benchmarking.[32] However, those results are not credible, because the model in column 1 is fatally under-specified.

[30] All models were estimated using our simpler Equation 5, but since Models 3 and 5 are strictly equivalent, our conclusions are unaffected when we estimate the original models.

[31] Column 1 corresponds to Table 1 Model 3 in *Benchmarking Across Borders*. Column 2 corresponds to Table 3 Model 2. Column 3 corresponds to Table 3 Model 5. Column 4 corresponds to Table 3 Model 8.

[32] Note that there is no evidence of benchmarking with respect to unemployment, even in this model.

Controls, lags, and fixed effects

Ensuring that results are robust to the inclusion of controls and a lagged dependent variable is a minimum standard for most modern research on economic voting. In Column 2 of Table 1, we follow KP and add the same control variables as in their article; column 3 includes the incumbent's vote share in the previous election; and column 4 includes both a lagged dependent variable and country fixed effects.

The three new models are consistent with conventional economic voting: The marginal effects of domestic growth in columns 2 to 4 are all positive and statistically significant. However, none of the three models supports benchmarking: The marginal effects of international growth in columns 2 to 4 are all indistinguishable from zero. As soon as we introduce control variables, a lagged dependent variable, or country fixed effects – widely recognized best practices in the field – the evidence of benchmarking evaporates.

Alternative measures of international growth

The models in Table 1 were all estimated using an index of international growth constructed by principal components analysis. This is KP's preferred measure of G_i, but the authors also consider two alternatives: A trade-weighted average of growth rates around the world, and the international median.

In *Benchmarking Across Borders*, the choice between those three measures is rather inconsequential, because KP conclude that the evidence supports benchmarking, regardless of how they measure G_i. Substantively, the authors take this to mean that "voters respond to their country's deviation from various measures of average international performance".[33] Moreover, KP do not offer a real theoretical defense of their preferred measure, and fit statistics do not give us strong reasons to favor one measure of G_i over another.[34]

Nevertheless, access to these two alternative measures of international growth is useful, because it allows us to probe the sensitivity of benchmarking tests to how we measure the reference point. In the online appendix, we replicate the eight regression models that KP estimated using aggregate-level data and their two alternative measures of international growth. None of those eight models shows evidence of benchmarking: The marginal effect of international growth is never distinguishable from zero.

Individual-level survey data

Moving beyond aggregate-level data, KP also study benchmarking using individual-level surveys. Once again, their empirical specification resembles Model 3, and the quantity of interest is the marginal effect of international growth. In the online appendix, we replicate all 12 of KP's individual-level models. None of those models allows us to reject the null of "no benchmarking."

[33] Kayser and Peress 2012, 669.

[34] Raftery (1995) considers that there is "strong" evidence against a model when its BIC is 6 to 10 points larger than another model. In Table 2 of their article, KP report that the gap in BIC between the principal components and the median growth models is between 3.2 and 4.4.

Three more empirical claims

In the online appendix, we consider three more empirical claims from the original article: (1) A statistically insignificant estimate of θ_i constitutes evidence of "full benchmarking"; (2) the substantive effect of decomposed growth is more important than the substantive effect of domestic growth; (3) at several points in time, the magnitude of the benchmarked economic vote is greater than the magnitude of the non-benchmarked economic vote. Our assessment is that these claims do not add credence to the theory.

Do voters benchmark economic performance?

In their article, KP "argue that previous research has fundamentally misunderstood and hence incorrectly estimated how economic assessments are made."[35] They contend that "voters respond more to national deviations from an international average rate of growth than to the growth rate itself."[36] They claim that their empirical analysis reveals "strong evidence of cross-national benchmarking on economic growth both at the aggregate and at the individual level, across time periods, and across subsamples."[37] Finally, after conducting extensive robustness checks, they conclude that their "main results are not altered"[38], and that the evidence is "clearly inconsistent with no benchmarking."[39]

We re-evaluated benchmarking on KP's own terms: Using their original data, logically equivalent statistical models, the same null hypothesis testing framework, and an evaluation criterion that they explicitly endorsed.[40] Yet, our substantive conclusions are strikingly different.

When models include control variables, a lagged dependent variable, or country fixed effects, we cannot reject the null of "no benchmarking." When we use alternative measures of international growth, we cannot reject the null of "no benchmarking." When we test the theory using individual-level survey data, we cannot reject the null of "no benchmarking." In fact, out of the 24 regression models that we replicated, only one model – without controls or lagged dependent variable – supports the theory. In the 23 other tests, the critical quantity of interest does not cross (or even approach) conventional thresholds of statistical significance.[41] Put simply, the evidence in *Benchmarking Across Borders* amounts to little more than a null result.

How to Test Benchmarking With Multiple Reference Points

The models considered above show little evidence of benchmarking. This surprising result could be an artefact of several factors. For instance, our models may be too simple to capture the complex processes at work, or KP's dataset may be too small to conduct well-powered tests. In this section, we consider how to adapt our barebones empirical framework to the more complex

[35] Kayser and Peress 2012, 680.

[36] 680.

[37] 662.

[38] 670.

[39] 669.

[40] Our hypothesis that the marginal effect of international growth should be negative is formally equivalent to the "partial benchmarking" hypothesis discussed in KP (2012, 668). According to KP, the appropriate way to test partial benchmarking is to check if a Wald test allows us to reject the null that θ_{y-i} and θ_i are equal. The p-value that this Wald test produces is exactly identical to the p-value of the δ_i coefficient in our model.

[41] The p-values of the δ_i coefficient for all 24 models are: 0.233, 0.006, 0.904, 0.472, 0.187 , 0.713 , 0.329, 0.316, 0.479, 0.454 , 0.575, 0.957, 0.330 , 0.209, 0.233, 0.690, 0.395 , 0.223, 0.702, 0.389, 0.246, 0.798, 0.443, 0.321.

case where voters compare domestic economic performance to multiple reference points. Then, we illustrate by studying a larger dataset drawn from a more recent study of benchmarking.

Aytaç[42] develops a reference point theory that is highly reminiscent of KP's, but which makes two important substantive changes. First, the author argues that voters use two reference points to assess their government's performance: The level of international growth (G_i) and their own country's historical level of growth (G_h).

Second, Aytaç points out that these reference points could be compared to two alternative measures of the incumbent's performance: Domestic growth during the election year (G_y), or domestic growth during the incumbent's full term in office (G_t). The term-based measure is preferable if we adopt a rational voter model, since such voters can extract more information about the quality of government by observing performance over a longer period. The election year measure is preferable if we take the view – dominant in political psychology – that voters are cognitively limited, myopic, and that they use end-heuristics when engaging in retrospective evaluations.[43] Here, we remain agnostic and estimate models using both measures.

In our framework, testing theories of benchmarking with multiple reference points is straightforward: We simply introduce the new reference point variable additively in Model 5. Again, there is evidence of benchmarking if the marginal effect of domestic growth is positive, and if the marginal effects of the benchmarks are negative.

In Table 2, we illustrate this by estimating six models using Aytaç's replication data.[44] In a first set of three models, we compare international and historical growth to domestic growth in the election year. In a second set of three models, we compare international and historical growth to the average domestic growth rate during the incumbent's full term in office. We include the same control variables as Aytaç.

All six of the models in Table 2 show evidence of conventional economic voting: The coefficients for domestic economic growth (G_y or G_t) are all positive and statistically significant. In contrast, none of the models allows us to reject the null hypothesis of "no international benchmarking": The G_i coefficient is never statistically significant at the $\alpha = 0.1$ level.

The two right-most models in Table 2 show evidence of historical benchmarking: The G_h is negative and statistically significant. However, it is important to point out that those models rely on a highly unconventional assumption. Indeed, they assume that voters have long enough memories to accurately compare the average level of growth during the incumbent's full term in office, to the average level of growth during the previous government's term. This assumption clashes with common wisdom in the field of economic voting, where "virtually all macro-studies assume a short lag, generally of one year".[45] Most studies of benchmarking also use short-term measures of domestic growth.[46]

In sum, the results in Table 2 offer strong support for the conventional theory of economic voting, but evidence of benchmarking is mixed. None of the models allows us to confidently reject the absence of international benchmarking, and the only models that support historical benchmarking require us to jettison the widespread assumption that voters are myopic.

[42] Aytaç 2018.

[43] Healy and Lenz 2014.

[44] In the online appendix, we explain why these models do not replicate Aytaç's faithfully.

[45] Lewis-Beck and Stegmaier 2013, 378.

[46] Ebeid and Rodden 2006; Kayser and Peress 2012; Powell and Whitten 1993.

TABLE 2: Benchmarking with multiple reference points and alternative measures of domestic growth. OLS regressions with country-clustered standard errors.

Domestic growth:	Election year			Term in office		
Reference point(s):	International	Historical	Both	International	Historical	Both
G_y	0.676***	0.732***	0.728***			
	(0.162)	(0.171)	(0.171)			
G_t				1.273***	1.387***	1.492***
				(0.254)	(0.257)	(0.273)
G_i	-0.002		0.032	-0.392		-0.410
	(0.273)		(0.268)	(0.291)		(0.288)
G_h		-0.308	-0.309		-0.567***	-0.572***
		(0.213)	(0.213)		(0.202)	(0.202)
Vote share lag	0.689***	0.691***	0.691***	0.673***	0.669***	0.673***
	(0.064)	(0.064)	(0.064)	(0.065)	(0.063)	(0.063)
Coalition	0.216	0.215	0.217	0.161	0.183	0.148
	(1.037)	(1.029)	(1.028)	(1.071)	(1.025)	(1.023)
ENP	-1.435***	-1.453***	-1.453***	-1.523***	-1.564***	-1.569***
	(0.366)	(0.359)	(0.359)	(0.393)	(0.377)	(0.385)
Presidential	-4.558***	-4.554***	-4.559***	-4.744***	-4.836***	-4.790***
	(1.108)	(1.113)	(1.113)	(1.120)	(1.093)	(1.100)
Re-run	12.284***	12.163***	12.165***	12.986***	12.852***	12.869***
	(2.485)	(2.467)	(2.468)	(2.426)	(2.360)	(2.387)
Constant	11.697***	12.587***	12.527***	11.853***	12.599***	13.362***
	(3.354)	(3.230)	(3.329)	(3.475)	(3.280)	(3.435)
Observations	460	460	460	460	460	460

Robust standard errors in parentheses

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

The regression models that we studied so far were relatively under-specified. Indeed, one of the major contributions of Powell & Whitten[47] was to point out that the level of economic voting depends on the institutional context (e.g., clarity of responsibility). Similarly, there are good reasons to think that benchmarking will vary across populations: some voters – such as those with high information – might engage in more relative economic evaluations than others.

If benchmarking is truly conditional, then the "pooled" models that we estimated above would be inappropriate, and our null results would not be surprising. For this reason, it is extremely important to develop regression models capable of testing conditional benchmarking hypotheses. Again, this is very easy to do in our simple empirical framework.

We use the same starting point as before (Figure 1): Benchmarking predicts that the marginal effect of domestic growth should be positive, and that the marginal effect of the reference point should be negative. If a moderating variable M increases (decreases) the salience of the reference point, then the marginal effect of domestic growth should be more (less) positive, and the marginal effect of the reference point should be more (less) negative where M is high.

This idea can be captured by a simple extension of Model 5:

$$V = \delta_y G_y + \delta_i G_i + \delta_{ym} G_y \times M + \delta_{im} G_i \times M + \delta_m M + \Gamma\Omega + \varepsilon, \tag{6}$$

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

As usual, a positive marginal effect of domestic growth ($\delta_y + \delta_{ym} M > 0$) would be consistent with both conventional economic voting and benchmarking. A negative marginal effect of international growth ($\delta_i + \delta_{im} M < 0$) would be consistent with benchmarking. The slopes of those marginal effects (δ_{ym} and δ_{im}) measure the extent to which M moderates relative economic assessments.

To illustrate how one can apply Model 6, we revisit a secondary set of tests from Aytaç (2018), where the author studies if benchmarking is more prevalent in countries with high trade intensity, GDP per capita, or average level of schooling.[48] We assess the moderating effect of all three variables,[49] include the same control variables as in Table 2, and use Aytaç's two alternative measures of domestic growth (G_y and G_t). The full regression results are reported in the online appendix.

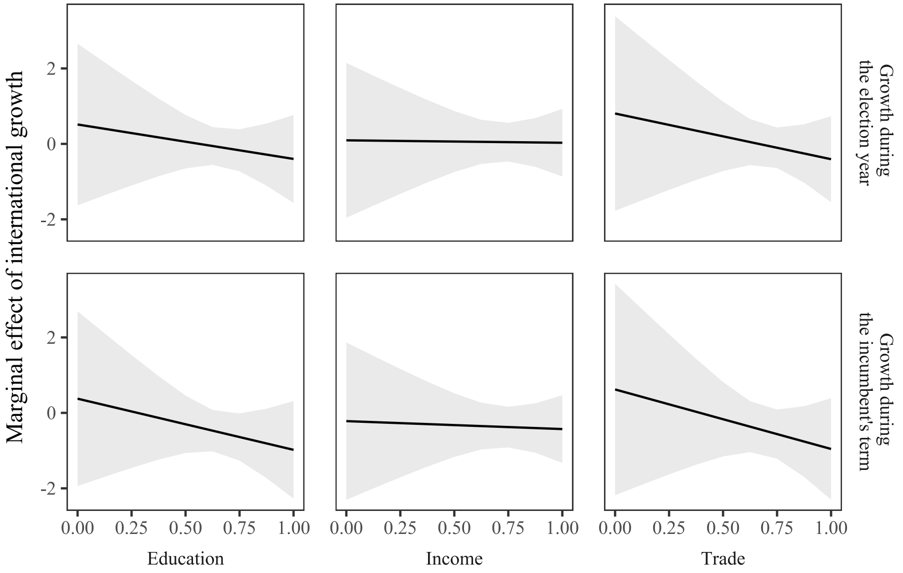
Figure 2 shows the estimated marginal effect of international growth in six models. None of the marginal effects is clearly negative, and most lines are nearly flat. These results, estimated using a dataset that is over twice the size of KP's, offer no evidence of international benchmarking, and no evidence that trade, income, or education increase the salience of comparative economic assessments.

[47] Powell and Whitten 1993.

[48] The interaction models that we report here are slightly different from those estimated by Aytaç (2018). In the online appendix, we take a close look at Aytaç's interaction specifications. Our discussion highlights some of the pitfalls of testing conditional benchmarking using composite variables and redundant regressors.

[49] For comparison, the moderators are all rescaled to the [0, 1] interval.

Figure 2: Marginal effect of international growth on votes for the incumbent in six regression models with three different moderators and two alternative measures of domestic growth. 95% confidence intervals in grey.



Conclusion

In this paper, we re-interpreted the theory of benchmarking and explained that, all else equal, it predicts that votes for the incumbent should be *positively* related to domestic growth, but *negatively* related to reference points. By re-casting the theory's predictions in terms of the marginal effects of domestic growth and the reference points, we showed that benchmarking could be tested using a simpler linear model which excludes duplicate regressors, immediately produces the relevant discriminating statistics, and greatly facilitates interpretation.

We re-analyzed data from prominent studies which have claimed to present evidence clearly supportive of benchmark. Across a range of models, we found robust evidence that domestic growth affects voting behavior, but very little sign of benchmarking. We therefore conclude that benchmarking is an interesting hypothesis, but that it is *not* supported by the available evidence.

These results should not be interpreted as a wholesale rejection of the benchmarking hypothesis. Indeed, it seems reasonable to expect that some populations may be more responsive to international or historical comparisons than others. For example, voters may be particularly attuned to the economic performance of neighboring countries or rivals.[50] At the individual level, some types of voters (e.g., politically sophisticated) may also be more prone to compare domestic with global economic performance or present economic conditions with previous ones. The idea that some voters evaluate economic performance in relative terms has some intuitive appeal, but the idea that most voters systematically and accurately compare with an "objective" benchmark seems rather implausible given citizens' cognitive limitations.

[50] Jérôme, Jérôme-Speziari, and Lewis-Beck 2001; Hansen, Olsen, and Bech 2015.

Perhaps most importantly, we have shown that there are great risks in using composite measures when it comes to testing theories of relative evaluation. We have demonstrated that there is a straightforward way to test the benchmarking hypothesis, which is to avoid composite variables, and to simply enter each term additively in the regression equation. With such an approach, we can formulate clear tests of the benchmarking hypotheses: all else equal, the benchmark should have a negative marginal effect on support for the incumbent party (or other relevant dependent variables). This simple empirical framework can also be extended in straightforward fashion to test theories of benchmarking with multiple reference points or context-conditionality. We hope to have provided clear guidelines for further research on this complex and important question.

Funding

We acknowledge the financial support of the Social Sciences and Humanities Research Council of Canada, and of the Fonds de Recherche Société et Culture du Québec.

References

Achen, Chris H., and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections do Not Produce Responsive Government.* Princeton: Princeton University Press.

Anderson, Christopher J. 2007. "The End of Economic Voting? Contingency Dilemmas and the Limits of Democratic Accountability." *Annual Review of Political Science* 10:271–296.

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

Bartels, Larry. 2012. "Elections in Hard Times." *Public Policy Research* 19 (1): 44–50.

Converse, Philip E. 2000. "Assessing the Capacity of Mass Electorates." *Annual Review of Political Science* 3:331–353.

Duch, Raymond M., and Randolph Stevenson. 2010. "The Global Economy, Competence, and the Economic Vote." *Journal of Politics* 72 (1): 105–123.

Ebeid, Michael, and Jonathan Rodden. 2006. "Economic Geography and Economic Voting: Evidence from the US States." *British Journal of Political Science* 36 (3): 527–547.

Fernàndez-Albertos, José. 2006. "Does Internationalisation Blur Responsibility? Economic Voting and Economic Openness in 15 European Countries." *West European Politics* 29 (3): 28–46.

Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections.* New Haven: Yale University Press.

Goplerud, Max, and Petra Schleiter. 2016. "An Index of Assembly Dissolution Powers." *Comparative Political Studies* 49 (4): 427–456.

Hansen, Kasper M., Asmus L. Olsen, and Michael Bech. 2015. "Cross-National Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic." *Political Behavior* 37 (4): 767–789.

Healy, Andrew, and Gabriel L. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58 (1): 31–47.

Healy, Andrew, and Neil Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of Political Science* 16:285–306.

Hellwig, Timothy, and David Samuels. 2014. "Voting in Open Economies: The Electoral Consequences of Globalization." *Comparative Political Studies* 40 (3): 283–306.

Jérôme, Bruno, Véronique Jérôme-Speziari, and Michael S. Lewis-Beck. 2001. "Évaluation économique et vote en France et en Allemagne." In *L'opinion européenne,* edited by Bruno Reynié D. et Cautrès. Paris: Presses de Sciences Po.

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.

Key, V.O. 1966. *The Responsible Electorate.* Cambridge (MA): Harvard University Press.

Leigh, Andrew. 2009. "Does the World Economy Swing National Elections?" *Oxford Bulletin of Economics and Statistics* 71 (2): 163–181.

Lewis-Beck, Michael S. 1988. *Economics and Elections: The Major Western Democracies.* Ann Arbor: University of Michigan Press.

Lewis-Beck, Michael S., and Mary Stegmaier. 2013. "The VP-function Revisited: A Survey of the Literature on Vote and Popularity Functions after over 40 Years." *Public Choice* 157 (3/4): 367–385.

Powell, Bingham G., and Guy D. Whitten. 1993. "A Cross-National Analysis of Economic Voting: Taking Account of the Political Context." *American Journal of Political Science* 37 (2): 391–414.

Przeworski, Adam, Susan C. Stokes, and Bernard Manin. 1999. *Democracy, Accountability, and Representation.* 2nd ed. Cambridge: Cambridge University Press.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–163.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–769.

Wolfers, Justin. 2002. *Are Voters Rational?: Evidence from Gubernatorial Elections.* Technical report 1730. Stanford: Graduate School of Business, Stanford University, March.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* Cambridge: Cambridge University Press.