1 Introduction

Among the most widespread regularities in the study of political behavior is the relationship between the state of the world and the vote. When times are good, incumbents tend to win reelection; when times are bad, sitting governments tend to pay the price at the polls. This phenomenon of retrospective voting – citizens registering their appraisals of past performance in their judgment of incumbents – motivates a vast empirical literature.[1] Past research finds voters around the world taking governments to task for outcomes in areas as diverse as economic growth, the prosecution of wars, and the making of prudent preparations for natural disasters. Yet it remains unclear if voters' judgments sensibly differentiate competent from incompetent incumbents. Notably, Achen and Bartels (2016) argue that voters' assessments reflect their idiosyncratic moods about the state of the world and that events beyond the incumbent's control – e.g., droughts or global financial crises – weigh heavily on these moods. Retrospective voting, in this view, is more akin to kicking the dog over a missed promotion than a mechanism for political accountability.

Debate about the normative significance of retrospective voting follows from uncertainty about how – and how well – voters manage what we term the "integration-appraisal" task. Retrospective voters of any stripe must reduce streams of performance information into a summary evaluation of incumbent competence. This necessitates the *integration* of information streams into a cognitively manageable impression of the state of things during an incumbent's term and the formation of an *appraisal* based on that impression. Theories of retrospective voting describe the process by which voters might – reliably or otherwise – manage integration and appraisal, but existing frameworks for studying retrospective voting are ill-equipped to test these arguments.

Numerous observational studies consider voters' capacity to meet various information-processing challenges inherent to retrospective voting (e.g., Quinn and Woolley, 2001; Ebeid and Rodden, 2006; Kayser and Peress, 2012; Stiers et al., 2020). However, because nature does not generally assign performance outcomes randomly, and because observational researchers typically cannot disentangle an incumbent's contribution to those outcomes from exogenous spillover, concerns about causal identification abound. This, of course, is the opening line of the experimentalist's sales pitch. A more tangible virtue of experimentation in the study of retrospective voting is the method's relative capacity to shed light on the integration-appraisal process. The ability to

[1] Retrospective voting is sometimes called "performance voting," and terms like "economic voting" refer to retrospective voting in a particular domain. See Healy and Malhotra (2013) for a review of the literature on retrospective voting.

manipulate the valence, variance, and source of performance outcomes – and the context within which those outcomes are observed – allows the experimentalist to study the psychology of retrospection in settings that are often difficult to observe in the "real world." In our view, experimental control of this nature is critical to advancing the study of retrospective voting.

Surprisingly, experimental research into the microfoundations of retrospection is quite limited. Those experiments that do exist generally simplify away one or both phases of integration and appraisal. In a typical design, researchers might expose participants to a static, "pre-integrated" summary of performance (e.g., GDP growth) and a "pre-appraised" interpretation thereof (e.g., "the rate of growth is good"). Thus, in both the typical observational and experimental studies, we may see that individuals respond to performance cues in judging an incumbent, but the core evaluative process remains largely unobserved.

In this Element, we explore the microfoundations of retrospective voting. We develop an abstract experimental framework that captures the streaming quality of information confronted by voters and *requires* them to arrive at their own appraisal of the information they encounter (Section 3). We present results from eleven experiments addressing foundational questions about the mechanisms driving the retrospective vote. We consider voters' basic capacities in assessing a variable stream of performance information (Section 4); how retrospective voters confront the challenge of interdependence, where spillover from extra-local forces obscures an incumbent's efforts (Section 5); and how voters seek out and use heuristics, that is, simplifying information about performance, in the integration-appraisal process (Section 6). Given our reliance on an abstract experimental design we also advance a case for abstract experimental approaches in the study of retrospective voting. In Section 2, we outline the logic that justifies such designs and explain how their unrealism facilitates, rather than impedes, the making of new generalizations.

1.1 Retrospective Voting As Integration and Appraisal

The literature contains scores of models of retrospective voting, some more explicit than others (e.g., Downs, 1957; Key, 1966; Fiorina, 1981; Ferejohn, 1986). For Berry and Howell (2007), retrospective voting is simply "the proposition that citizens examine whether the state of the world has improved under a politician's watch and vote accordingly" (p. 844). In their landmark study of economic voting, Duch and Stevenson (2008) proceed formally, presenting a model that rests on a specific decomposition of the variance of economic growth. The unifying thread, and the foundation of our account, is the idea that voters' evaluations of the state of the world motivate their decision to

reelect or replace an elected official. This implies that voters acquire information about prevailing conditions and use at least some of it to form judgments of their government's performance. As those judgments improve, so too does the likelihood that the voter supports reelection.

This process unfolds in an environment where voters encounter streams of performance cues – e.g., quarterly jobs reports or news stories about crime – across the timeline of electoral politics. Whether these flows are sporadic, uneven, or confounded by external spillover, acknowledging the streaming quality of this information highlights the necessity for retrospective voters to undertake some sequence of integration and appraisal. Before judging an incumbent, retrospective voters must *integrate* some of the information they encounter into a summary impression. Consciously or otherwise, this requires weighing information of varying relevance and quality. Voters must also render an *appraisal* of their impressions and ultimately apply that to the vote decision (cf. Downs, 1957, pp. 45–46). Consider a voter concerned with the price of groceries in their state. Over time, they are exposed – via direct experience and mediated reports – to information they may recognize as bearing on food prices. How do they weigh and combine these diverse indicators to form an overall summary of affordability during their governor's administration? How do they decide whether that summary indicates good or bad performance?

Our approach to retrospective voting through the framework of integration and appraisal is significant for several reasons. First, because the task is fundamental to retrospective voting, the integration-appraisal framework encompasses a wide range of more specific behavioral-psychological and informational assumptions. The task obtains whether voters' appraisals of performance are myopic (e.g., Healy and Lenz, 2014) or based on integration over longer time horizons (e.g., Stiers et al., 2020). It allows that voters may be relatively discerning (e.g., Hellwig, 2001) or indiscriminate (e.g., Achen and Bartels, 2016) in identifying their incumbent's contributions to performance outcomes. The integration-appraisal framework encompasses both the assumption that voters seek to *sanction* their governors for past outcomes – to incentivize good performance (Key, 1966) – and the possibility that voters use past outcomes to infer incumbent competence, thus informing the *selection* of quality rulers (Duch and Stevenson, 2008). And the task holds whether voters encounter performance information directly or that information is mediated by friends, television, or Twitter. Performance necessarily unfolds over time, even if its level is static. Whether relying on direct experience or the reports of others, and whether information arrives as "raw data" (e.g., unemployment is 9 percent) or suggests an evaluative interpretation (e.g., "Have you heard how high unemployment is?"), integration across multiple pieces of performance information is required.

Second, unpacking the integration-appraisal task clarifies the problematic nature of appraising performance, even when doing so involves only one piece of information. A Canadian voter suffering respiratory symptoms must wait a week to see their family doctor. Is the public health care system performing well or poorly? Answering such questions requires an account of how voters judge the quality of a given level of observed performance.

Third, the framework highlights the difficulty of forming a summary impression from multiple pieces of information. The need for voters to integrate across streams of variable performance increases the potential burden of retrospective voting relative to a world – like the one found in many experimental tests of retrospective voting – that simply reveals pure signals of incumbent competence. Some strategy must be applied, consciously or otherwise, to weigh and reduce the multiplicity of performance cues into a manageable impression. This challenge is aggravated in an era of globalization, where performance *here* is increasingly tied to actors and actions *over there*. This problem of interdependence, or spillover, makes it more difficult for voters to identify their incumbent's efforts. To be sure, the "rule" voters apply in confronting this challenge may be normatively sound (e.g., systematic weighing of discrete performances) or quite arbitrary (e.g., attending only to whatever information is cognitively available when making a judgment). Our argument is simply that the availability of multiple cues, and the confounding of those cues by spillover from multiple actors, is a generic feature of retrospective voting.

In conceptualizing performance evaluation in terms of the integration-appraisal task, we emphasize that we are not advancing a novel theory of retrospective voting. On the contrary, we regard the assumption that voters observe, combine, and evaluate indicators of performance as an essential feature of *all* retrospective voting theories, including those that assume voters manage the task poorly. With the experiments reported in this Element, we evaluate a family of specific theories that, despite their varying assumptions, share a general premise: that retention of incumbents is related to their performance *via* a process of integration and appraisal of performance cues.

In summary, retrospective voting theory implies that voters encounter information bearing on the state of the world that they use to judge government performance.[2] An important feature of that information is that it generally manifests as a stream of cues. The retrospective voter somehow utilizes the information in the stream to arrive at an impression of performance and forms

[2] Discussing the integration-appraisal task as a linear, two-step sequence is only an expository convenience. Alternatively, voters may first appraise incoming bits of information and then integrate over appraisals before rendering a final appraisal (see Section 4.3.2).

an appraisal that they apply come Election Day. The key point is that both integration and appraisal are inherent to this process.

1.2 Prior Experiments in Retrospective Voting

In view of the vastness of the retrospective voting literature, experimental tests are surprisingly uncommon and recent. A systematic search of the political science literature uncovers just thirty-four published articles containing experimental tests of retrospective voting.[3] All but four of these were published after 2010, and roughly a third do not involve questions that require the manipulation of performance evaluations in any way.[4]

Studies that do explicitly "treat" performance evaluations generally short-circuit the integration-appraisal task in whole or in part, with subjects encountering information that requires no integration and/or no appraisal. Treatments often combine high-level verbal summaries of performance outcomes in a real or hypothetical setting with an explicit or implicit appraisal. For example, in Tilley and Hobolt's (2011) study of attributions of responsibility in retrospective voting, participants in one condition read: "Experts say that not only have economic conditions deteriorated a lot over the last year, but the British economy is doing considerably worse than most other countries" (p. 323). The statement provides an integrated summary of economic performance; it also conveys an implicit appraisal through its evaluative tone ("conditions deteriorated a lot") and the unfavorable comparisons to other economies (see also Sigelman et al., 1991; Klasnja and Tucker, 2013; Malhotra and Margalit, 2014; Simonovitz, 2015).

Several experiments provide similar pre-integrated summaries of performance outcomes but without cues to guide participants' appraisal. Clinton and Grissom (2015), for example, examine the impact of revealed performance on evaluation of educational institutions in Tennessee. Their treatments inform participants about the percentages of students "performing at grade level or better on Tennessee's end-of-grade math tests" (p. 365). This allows participants to form their own appraisals of competence even as it obviates the need for voters to reduce streams of performance cues into a summary of their own.

Finally, some experiments treat performance evaluations with information that is not summarized but nonetheless provide a weak basis for conclusions

[3] Using the Web of Science database, in February 2022 we queried political science journal abstracts for intersections of "experiment" with "retrospective voting," "economic voting," or any of the various related word combinations (e.g., "economic evaluation"). We augmented the results with pieces of which we were already aware. We reviewed the ~150 abstracts to determine which involved experimental work on retrospective voting.

[4] Some seek instead to manipulate the *salience* of performance evaluations to examine the existence of retrospective voting rather than its microfoundations per se.

regarding citizens' handling of the integration-appraisal task. For example, Bhandari et al.'s (2023) field experimental study of the impact of benchmarks on retrospective voting treats participants with information regarding legislator performance along five separate dimensions. Yet tests of the impact of the performance information rely on comparisons with a control condition in which no performance information was presented. As such, it is unclear whether the treatment effect reflects the acquisition – and integration – of new performance information or simply the priming of performance considerations.

None of this is meant to suggest that these studies fail to address the research questions on which they focus. On the contrary, these studies offer key insights in the retrospective voting literature. Our point is that those insights rely on designs that cannot speak directly to voters' capacity to integrate across multiple pieces of performance information and to form appraisals that are unaided by explicit or implicit cues. In this sense, the far-reaching literature on retrospective voting rests on assumptions about an integration-appraisal task that has rarely been studied in a controlled setting.

An important exception is Huber et al.'s (2012) study of bias in retrospective evaluation. The logic of the design is to study performance evaluation in a highly simplified and abstract setting that "allows [them] to understand whether suboptimal decision making is a function of the complexity of the real world . . . or is instead due to basic limitations in individuals' retrospective abilities" (p. 721). Participants in these studies play an incentivized game that centers on the decision to replace or retain an "allocator." The allocator provides tokens as payments that subjects later redeem for cash. The allocator's underlying competence, or type, is drawn randomly from a known distribution; the payment in each period is drawn from a distribution centered on the allocator's type. Participants observe a sequence of sixteen payments before electing to retain their allocator for another sixteen payments or draw another at random from a known distribution.

From our perspective, the key feature of Huber and colleagues' design is that performance information is conveyed as a stream of discrete outcomes, unaccompanied by an explicit or implicit appraisal.[5] Participants must, therefore, form an impression on their own. Having formed such an impression, the task of judging whether the allocator clears the bar for retention is also left to participants. In a way that is seemingly unique in the literature, the incentivized

[5] As we discuss in Section 3, the game's instructions allow participants to infer an "optimal decision rule for risk-neutral participants" (Huber et al., 2012, p. 723), which then implies an appraisal of average performance. In our view, this feature of the design leaves considerable room for participants to arrive at independent appraisals of performance, though we expect appraisals to be informed by the instructions (see the discussion in Section 4.1.1).

allocator design leaves the central information-processing task of retrospective voting entirely to the participant.[6]

1.3 Our Experimental Approach

The experimental studies in this Element build on – and go considerably beyond – Huber and colleagues' basic framework to address new questions and long-standing debates about retrospective voting (notably, doing so leads us to draw quite different conclusions; see, especially, Section 7.1). Here, we outline our approach and describe our experiments in general terms, leaving the details to later sections. For reference, Table 1 summarizes the eleven experiments we present in this Element.

Table 1 Summary of retrospective voting experiments

Experiment	Performance streams (n)	Motivating questions
1	Single (248)	Given a stream of variable performance output, can participants discriminate between competent and incompetent incumbents?
2	Single (403)	Does performance variability inhibit retrospective voting?
3	Single (719)	Negativity bias: Are participants more inclined to punish poor performance than to reward good performance?
4	Single (764)	Recency bias: Do participants place more weight on performance late in an incumbent's term?
5	Multiple (849)	How do participants judge an incumbent whose performance is confounded by a high- versus low-variance disturbance?
6	Multiple (368)	Can participants assess an unobstructed performance signal without reference to an exogenous disturbance?

[6] Huber et al. (2012) are largely silent on the theoretical significance of their unique design in this regard, though they recognize their design requires participants to "infer" performance from observed "payouts" (p. 720).

Table 1 (cont.)

Experiment	Performance streams (n)	Motivating questions
7	Multiple (550)	How do participants judge an incumbent whose performance is obscured by a complex disturbance?
8	Multiple (720)	Given a choice among summary measures of performance, do participants actively seek "pre-benchmarked" indicators?
9	Multiple (1,376)	Does exposure to summary reports of incumbent competence mitigate the benchmarking response?
10	Single (394)	Given a choice among benchmarks of varying quality, do participants select good benchmarks?
11	Multiple (918)	Do participants adjust their comparisons for "shocked" benchmarks?

Our framework asks participants to serve as factory managers observing the performance of a new worker (or workers). After sixteen weeks of production, the participant must elect to extend their "incumbent" worker's contract or to hire a replacement. The incumbent's competence is unknown at the outset, the stream of production is variable, and information regarding incumbent performance is sometimes confounded or obscured by the presence of co-workers; the game ends with a choice of consequence between the incumbent and a lesser-known replacement. We attempt to motivate participants to maximize worker performance by assigning cash bonuses in proportion to the number of units their workers produce. We do not suggest an appraisal of worker performance, except when doing so is necessary to answer a question of interest (i.e., in Exps. 3 and 11).

The principal virtue of the factory-worker metaphor, relative to Huber et al.'s (2012) allocator, is that we can draw on stereotyped knowledge of workers and factories to simplify exposition of the features of our experimental environments. This gives us the flexibility to address a wide range of fundamental questions in the retrospective voting literature with only minor changes to game design and task instructions. The experiments in Section 4 (Exps. 1–4), for example, all consider an individual's capacity to integrate and appraise over a

single stream of incumbent performance information. With slight adjustments to the presentation of information or the parameters of worker performance, we can test a range of hypotheses, concerning, for example, recency bias (Healy and Lenz, 2014; Stiers et al., 2020) and negativity bias (Baumeister et al., 2001).

The flexibility of our experimental framework – and the familiarity of the factory analogy – also enables us to create more complex tasks without short-circuiting the integration-appraisal process. This is critical as we explore the challenge of retrospection under interdependence. How individuals evaluate government performance when external spillover – the coming together of endogenous and exogenous streams of performance – obscures the incumbent's efforts is a central focus in recent observational research on retrospective voting. Yet we know of no experiment that asks participants to integrate and appraise over multiple and/or confounded streams of information. Section 5's experiments (Exps. 5–8) introduce the performance of peer or "comparator" workers, some of whom arrive late or depart early, to obscure the incumbent's efforts. In so doing, these experiments permit diagnostic tests of three conventional models of retrospective voting under interdependence: rational discounting (Duch and Stevenson, 2008), benchmarking (Kayser and Peress, 2012), and blind retrospection (Achen and Bartels, 2016).[7]

Finally, the experiments in Section 6 (Exps. 8–11) turn to the use of heuristics – or simplifying information about performance – in the integration-appraisal process. With only minor changes to our basic games, we study individual preferences for different heuristics and the extent to which the availability of objectively "good" performance indicators shapes the use of particular heuristics in the evaluation process.

1.4 Partisan Identity and Experimental Tests of Retrospective Voting

One of the most common questions we received as we presented findings from this project concerns the role of partisan identity in retrospective voting. Given what we know about the effect of partisanship on political behavior, should that variable not play an explicit role in our otherwise abstract framework? Put differently, shouldn't we abandon the factory worker in favor of a more politically "realistic" design? Our answer – which is no – goes to the heart of our theoretical and empirical approach. While Sections 2 and 3 elaborate our understanding of these issues in detail, given the centrality of partisan social identity to contemporary scholarship, we briefly summarize our perspective here.

[7] Our discussion of Experiments 5–9 reproduces parts of Hart and Matthews (2022).

The motivation for this Element is theory-testing: to evaluate and advance models of retrospective voting. We aim to generalize from the behaviors we observe in the experiments to the theories themselves and not, for example, to the incumbent's vote share in a specific election or under specific economic conditions. Introducing variables extraneous to the models we test defeats the purpose of experimentation and, for theory-testing endeavors, limits generalizability. As we argue in Section 2, efforts to increase the "mundane realism" of experimental treatments and settings does not confer external validity on theory-testing experiments (see also McDermott, 2002; Druckman, 2022). Similarly, the omission of concrete realities beyond the scope of the theories being tested does not threaten external validity. When the goal is theoretical rather than statistical generalization, an experiment is externally valid "if it is an instance of the theory it tests and neither adds nor subtracts anything from the theory" (Zelditch, 2007, p. 96). This approach, though not novel (e.g., Lupia and McCubbins, 1998), highlights a major point of departure from much contemporary thinking in experimental political science about generalizability.

Consider the possibility that partisanship moderates the retrospective vote. Would assigning a salient social identity to the incumbent "workers" in our experiments moderate responses to their performance? Probably so. For instance, Brutger et al. (2022) find that adding details and context to experimental settings may dampen treatment effects. But so too, we think, would presenting the instructions – to our English-speaking respondents – in Dutch. Keeping in mind that neither variable is an essential element of models of retrospective voting, would introducing these extraneous moderators increase our leverage over the theories we test? No. We would join the queue of observational and experimental studies that we have already suggested fail to identify and effectively explore the integration-appraisal task. To be sure, theoretical claims regarding party identification's impact on how voters seek out, integrate, and appraise performance information are well worth testing. Yet these processes are extraneous to the models we test, and exploratory mediation analysis must play second fiddle to foundational tests of retrospection. After all, we have seventy-odd years of empirical study and few clear insights into how, and how well, voters manage the central task of retrospective voting.

1.5 Preview of Findings

Our studies reveal a retrospective voter who can – and often does – make sensible use of the performance information they encounter. Participants in our experiments reward and punish incumbents in accordance with their productivity and do so without the aid of simplifying summaries and evaluative interpretations.