Unmasking accountability: Judging performance in an interdependent world*

Austin Hart† American University

J. Scott Matthews ‡ Memorial University

November 6, 2019

Abstract

As the conditions in one jurisdiction increasingly reflect actions and events in others, signals of incumbent performance grow more obscure. Yet we know relatively little about how voters judge incumbents in the context of interdependence. We use a series of simulated voting tasks to examine three theoretical possibilities: signal extraction, blind retrospection, and benchmarking. Across three experiments where we obscure incumbent competence, we find consistent evidence of benchmarking — subjects rewarded incumbents, capable or otherwise, who outperformed a peer. Benchmarking was evident in information processing, information seeking, and in both hard and easy tasks. Our findings have implications for the study of performance voting under globalization and similar conditions of interdependence, and raise questions regarding the availability of performance information across different domains of government action.

*Earlier versions of this paper were presented at the Annual Conference of the American Political Science Association as well as faculty seminars at American University, Memorial University, and the University of Kansas. Special thanks to Mark Kayser and Dave Peterson for their invaluable feedback. Matthews acknowledges the generous support of the Alexander von Humboldt Foundation and the Mannheim Centre for European Social Research at the University of Mannheim.

†School of International Service, American University. Email: ahart@american.edu

‡Department of Political Science, Memorial University. Email: scott.matthews@mun.ca

Introduction

What do changes in the rate of unemployment or violent crime tell us about government competence? Students of voting behavior have long assumed that citizens readily attribute responsibility for the state of things to elected officials (e.g. Kramer 1971; Lewis-Beck 1988), and evidence of performance voting abounds.[1] Most notably, a growing economy appears to buoy the election prospects of incumbent-party candidates.[2] However, the challenge of judging a representative's stewardship is aggravated by the growing linkages between individuals, businesses, and governments around the world. These interactions—globalization in the broadest sense—yoke performance outcomes in one constituency to actions and events elsewhere, constraining government agency on the one hand and creating uncertainty about government competence on the other.[3] After all, voters in an interdependent system experience changes in wellbeing, but they cannot observe their representatives' contribution. How, if at all, do voters in these contexts unmask an incumbent's competence?

Prior research highlights three competing theories of performance voting under interdependence: blind retrospection, signal extraction, and benchmarking. Blind retrospection (e.g. Achen and Bartels 2016) posits that voters ignore the problem of interdependence altogether. Rather than parse "contaminated" streams of performance data, the blindly retrospective voter sees any change in wellbeing as an indicator of the government's competence. By contrast, the signal extraction thesis (e.g. Duch and Stevenson 2008) contends that voters judge governments with a high degree of discernment, extracting competence signals by applying a discount factor. Where external forces come to weigh heavily on local conditions, voters struggle to find clear evidence of their representatives' contributions. So they learn to discount performance outcomes and credit the government for only a fraction of observed changes. Finally, benchmarking (e.g. Kayser and Peress 2012), holds

[1] Prior studies report that citizens hold elected officials accountable on a wide range of performance outcomes, including corruption (e.g Ferraz and Finan 2008), economic growth (e.g. Kiewiet 1983), national security (e.g. Ladd 2007), and disaster relief (e.g. Healy and Malhotra 2010). Attribution of responsibility appears to be dynamic and issue-specific (Larsen 2018), but the magnitude of the response varies by political context (e.g. Tavits 2007; Hobolt, Tilly, and Banducci 2013). For a review of the literature on retrospective voting, see Healy and Malhotra (2013).

[2] That citizens hold governments accountable for the state of the economy is well documented (e.g. Fiorina 1981; Kinder and Kiewiet 1981; Simonovits 2015; Lenz 2012). However, the extent to which voters turn to the economy at the ballot box varies with institutional context (e.g. Powell and Whitten 1993) and campaign dynamics (e.g. Hart 2016).

[3] There is a rich literature on globalization and its economic and political impacts. Of direct concern to performance voting, prior research suggests that globalization changes how parties compete (e.g. Hellwig 2015), the policies elected representatives pursue in office (e.g. Case, Hines, and Rosen 1993), and how voters respond (e.g. Hellwig 2008).

that voters cope with interdependence by judging governments in relative terms. Here voters see competence not in absolute levels of performance but in the gaps between conditions at home and conditions among external peers. If, for instance, a state in a wheat-exporting region is growing at 4% while neighboring states are growing at 6%, benchmarking voters see the gap—a two-point *contraction*—as the government's contribution. In sharp contrast, the blindly retrospective voter credits the government for the full 4% expansion, and signal extractors credit government for some fraction of the same (between 0-4 points).

Blind retrospection, signal extraction, and benchmarking offer very different visions of performance voting and, consequently, the accountability function of elections. Each approach has support and challenges in empirical research, almost exclusively in observational studies. Much of the confusion, we argue, emerges because these approaches rest on competing but largely untested assumptions about how voters evaluate performance in the context of interdependent outcomes. In this paper, we use a series of simulated voting tasks to test these competing arguments directly. In a series of three experiments, we ask participants to judge the performance of a worker in a simulated factory and then "vote" to retain or replace this "incumbent." As with a newly elected official, the worker's level of competence is unknown and his/her output is both variable and obscured by external forces—in our voting tasks, these forces reflect the efforts of peers. The objective is to select workers who optimize the factory's total production—a quantity we link to a material payout.

The first experiment assesses individuals' capacity to extract a performance signal obscured by a high- versus low-variance comparator (akin to assessing economic performance in a relatively open versus closed economy). The second tests participants' ability to evaluate a clear—unobscured—performance signal without reference to the efforts of an irrelevant peer. Our third study assesses individuals' ability to extract performance signals obscured by the efforts of multiple comparators.

We report several important findings. Key among them is the consistent evidence of a benchmarking response that may help explain why prior studies offer contradictory support for each of the approaches considered here. In fact, we observe patterns of behavior consistent with elements of blind retrospection—participants held their worker responsible for aggregate factory production and exogenous shocks—and the signal extraction theses—subjects successfully extracted signals of worker competence. Notably, we find that the magnitude of the performance vote declines in the face of high-variance exogenous shocks (i.e., the other workers' performance). Yet, underneath these

responses is clear evidence of benchmarking. Rather than discard the irrelevant peer's performance as extraction theory predicts or overreact to it as blind retrospection implies, our participants used it as a baseline against which to judge the incumbent worker. Participants in both hard and easy tasks rewarded incumbents—competent or otherwise—for outperforming their comparators. We also observe a clear preference for pre-benchmarked indicators of performance over direct measures of competence.

We begin by detailing the competing theories of performance voting under interdependence and outlining our experimental design. We then proceed by study, highlighting the unique elements of each task and presenting results. The conclusion synthesizes the findings, relates them to prior research, and outlines implications for democratic accountability and possible avenues for future research.

Confronting Interdependence: Competing Approaches

Evaluating government performance in the context of interdependence is a fundamental challenge of representative democracy, and it confounds voting over a wide range of issues and across levels of government. Success in reducing a city's gun violence, for instance, depends in part on gun control and law enforcement efforts in other jurisdictions (e.g. Cook and Braga 2001; Pierce et al. 2004). Voters in these contexts confront the challenge of judging outcomes driven by the efforts of numerous actors, only some of whom voters are in a position to discipline.[4] The failure to decompose such outcomes in a way that allows voters to judge incumbents for "their" part in the process disrupts the accountability function of elections. In the worst-case scenario, an incumbent whose "own" performance is above average fails to be reelected owing to spillover from below-average performance in other jurisdictions.

Research in political science and economics suggests three ways in which voters judge performance for interdependent outcomes: signal extraction, blind retrospection, and benchmarking.

[4]Note the distinction between the problem of interdependent outcomes and collective outcomes. While both are multiagent processes in which the agents' individual efforts are unknown to the principal, agents in collectives work together—typically with varying degrees of influence—towards a joint outcome (e.g. the parties of a governing coalition working to combat unemployment). Moreover, voters typically have the power to discipline every member of the collective. The challenge is resolving the uncertainty that surrounds each agent's power within the group. By contrast, interdependence—the focus of our study—is the contingency of one agent's or one collective's performance on the performance of agents or collectives in constituencies beyond the voter's reach. The challenge is in identifying the agent's performance independent of the spillover.

In the *signal extraction* account, citizens manage the uncertainty of interdependence with a high degree of discernment, capably parsing "contaminated" performance shocks and extracting the true signal of government competence. Notably, Duch and Stevenson (2008; also: Alesina and Rosenthal 1995) argue that instrumentally rational voters isolate the government's performance by discounting economic movement[5] according to the ratio of the variance in exogenous shocks (spillover or luck) to competence shocks (government-driven fluctuations). Interdependence amplifies the noise and inhibits voters from seeing the government's efforts underneath broader trends in wellbeing. So voters learn to discount these trends, attributing more and more to external forces rather than government competence.

Prior studies show that voters behave "as if" they process performance data as signal extraction theory predicts[6]. In particular, voters appear to be aware of unexpected macroeconomic shocks (e.g. Duch and Stevenson 2010) and, in open, interdependent economies, are less likely to report that elected officials hold sway over economic outcomes (e.g. Alcañiz and Hellwig 2011; Hellwig 2015). Voters also appear less likely to hold governments accountable for national economic conditions in open economies (e.g. Duch and Stevenson 2008). Furthermore, individuals capably discriminate between local and national performance (Ebeid and Rodden 2006), and judgments of responsibility adapt as policymaking authority is passed to lower levels of government (e.g. Larsen 2018; León 2012).[7] These studies show, among other things, that citizens possess reasonable perceptions of the degree to which overall wellbeing reflects exogenous forces. This research does not, however, demonstrate that individuals use such information to pass judgment on elected officials; that is, the studies provide no test of the key mechanism: discounting by the variance of exogenous shocks. This gap in the literature is a concern given research on common biases in performance evaluations.

One alternative to the signal extraction logic is *blind retrospection* (Achen and Bartels 2016). In this view, voters never confront the problem of interdependent performance in any meaningful sense. Instead, they reward and punish incumbents for aggregate performance, including outcomes

[5] That is, unexpected variance that does not reflect the natural rate of economic growth or prior decisions of electorally-dependent actors (Duch and Stevenson 2008, pp. 141).

[6] A closely related, but theoretically distinct, line of research focuses on "clarity of responsibility". Work in this area highlights how political institutions blur or sharpen lines of responsibility for collective outcomes. Divided government, for instance, may blur the lines of responsibility and allow representatives to avoid sanction (e.g. Powell 2000; Samuels 2004; Anderson 2007; alternatively: Norpoth 2001; Carlin, Love, and Martínez-Gallardo 2015).

[7] Note, however, that Alesina and Rosenthal (1995) find no evidence that voters respond to variation in the signal-to-noise ratio.

over which political authorities exercise no control. Prior studies find, for instance, that voters take authorities to task for droughts (e.g. Cole, Healy, and Werker 2012), sporting victories (e.g. Healy, Malhotra, and Mo 2010; Healy and Malhotra 2010), shark attacks (e.g. Achen and Bartels 2016), and, contrary to the extraction logic, market shocks from other countries (e.g. Campello and Zucco 2016; Leigh and McLeish 2009; Hayes, Imai, and Shelton 2015) and other levels of government (e.g. Wolfers 2007; Gelineau and Remmer 2006).[8] Moreover, judgments of performance within the government's sphere of control are seen to be myopic (e.g. Healy and Lenz 2014) and biased by exogenous performance shocks (Huber, Hill, and Lenz 2012).

In stark contrast to sophisticated signal extractors, blindly retrospective voters do not parse responsibility for changes in prevailing conditions. In hard times, the blindly retrospective voter is likely to lash out at incumbents "whether or not objective observers can find a rational basis for blame" (Achen and Bartels 2016, pp. 118). In the worst-case scenario, owing to cognitive limitations or the effects of emotion on decision-making, voters "hold incumbents accountable for uncontrolled events even when they believe the incumbent is not responsible (for the event or its correction) and even when they receive distinct signals" (Huber, Hill, and Lenz 2012, pp. 731). These results suggest that voters are either unable or unwilling to process performance information in the way that extraction theory posits: discounting observed outcomes by the ratio of exogenous to endogenous variation. Rather, blindly retrospective voters hold incumbents to account for outcomes in a given domain irrespective of the incumbent's hand in creating them.

A second alternative to signal extraction is *benchmarking*.[9] Benchmarking voters resolve the uncertainty surrounding their government's performance by way of comparison an external peer or reference group. If states in an interdependent network are subject to the same pressures, then state-level deviations from the common trend become the key indicator of government stewardship. What matters is not performance per se. Rather, voters reward incumbents who outperform their peers and punish those who underperform.

[8]Note that, though often cited, research on the response to shark attacks, floods, and sporting victories may represent false positives (e.g. Fowler and Montagnes 2015; Fowler and Hall 2018) or is, at least, inconsistent (e.g. Bodet, Thomas, and Tessier 2016).

[9]Benchmarking is sometimes referred to as "relative performance voting" (e.g. Aytac 2017) or, in economics, a "yardstick" response(e.g. Holmstrom 1982; Besley and Case 1995; Hansen, Olsen, and Bech 2015). We use the term "benchmarking" throughout the paper for consistency.

Consistent with the benchmarking thesis, prior studies show that voters around the world evaluate national wellbeing relative to regional and international averages (e.g. Kayser and Peress 2012; Aytac 2017) and, in the United States, judge state economic performance against national outcomes (Wolfers 2007).[10] Most importantly, benchmarking may explain the same cross-national and over-time variation in performance voting identified in the signal extraction literature. To the extent that this is the case, the observed relationship between variance in exogenous economic shocks and the extent of performance voting (e.g. Duch and Stevenson 2008) reflects a spurious correlation: the fact that "[t]he larger international component of an open economy leads to smaller deviations from [the global benchmark] and, hence, a weaker economic vote" (Kayser and Peress 2012, pp. 664). However, recent work raises serious questions about these observational findings (Arel-Bundock, Blais, and Dassonneville 2018).[11]

Relative to extraction, benchmarking imposes a lower "calculative" burden on voters, requiring a simple comparison between "local" performance and that of a suitable benchmark. When the members of the chosen peer group are subject to the same dependencies, benchmarking may even represent the optimal mode of uncovering the individual contribution to performance (Holmstrom 1982). Survey experiments show that individuals primed with pre-benchmarked comparisons engage in benchmark evaluations (Hansen, Olsen, and Bech 2015), and observational studies highlight pre-benchmarked economic data in news reports as one plausible driver of comparative evaluations (Kayser and Peress 2012).

Experimental Design

The logic of our experimental task is simple. We present participants with a stream of irrelevant (exogenous) performance data and then, in a second stage, introduce relevant performance data to the stream. How and when participants encounter these streams differs in each study, but all share the same skeleton. Here we describe the structure of the common task and our two core designs – the "signal extraction" and "signal recognition" designs. Details of how each experiment builds on

[10] There is also evidence of policy benchmarking—officials adjusting the tax rate, for instance, based on the tax rate adopted in neighboring jurisdictions (e.g. Besley and Case 1995). Though benchmarking competition across governments may be one mechanism of policy diffusion (e.g. Shipan and Volden 2008; Tiebout 1956), policy benchmarking is a distinct phenomenon. Note, however, that the focus of the current study is restricted to performance benchmarking on the part of voters.

[11] See, however, the response from Kayser and Peress (2019).

these core designs are left to later sections. See the Supplemental Materials for question wording, instructions, tests of rules comprehension, randomization and robustness checks, and participant demographics.

For each experiment, we recruited U.S. adults to play a short decision making "game" using Amazon.com's Mechanical Turk (MTurk) marketplace.[12] We offered a baseline payment of $0.30 with the chance to earn a bonus averaging $0.67 for good gameplay. From August 9-September 13, 2018, 2,674 individuals accepted the task. To mitigate temporal bias in the available worker pool (Casey et al. 2017), we automated calls for batches of 50 to 100 respondents at fixed intervals over many days. After consenting to participate in the study, subjects report basic demographic information and answer a screening question.[13] The screen requires that subjects provide a non-obvious answer to a simple question. Subjects who do not read the question carefully, or who simply "click through," are likely to select an incorrect response ($Pr = 0.83$). We removed 23% of potential respondents for failing the attention screen.[14] The 1,950 who completed the task did so in under seven minutes on average.

Our core experiment is an applied extension of Huber, Hill, and Lenz's (2012) incentivized allocator game. We ask participants to serve as factory supervisors overseeing the efforts of new employees, a "Target" worker – akin to the incumbent candidate or party – and a "Comparator." The task is to observe the number of units their workers produce over a sixteen-week period and then, based on performance, vote to extend their Target's contract for another sixteen weeks or hire a replacement. However, and for different reasons in each experiment, the Target does not begin work at the factory until week 9. Limiting the vote to the Target worker renders information about the Comparator's performance exogenous and is analogous to the political decisions voters make in the real world. For instance, German voters might observe economic performance in both Germany and France, but they only get the chance to vote in Germany. The Target worker's delayed start, then, gives subjects eight weeks, or rounds, to identify the parameters of the disturbance before the relevant performance data enters the production stream. The goal in each study is retrospective accountability: select the most competent workers. Participants earn a cash bonus proportional to

[12] We programmed the task in Qualtrics. MTurk workers who accepted the task linked to the study and then entered their completion code to verify their work.

[13] The respondents in our study were 45% female and 76% white with a median age between 30 and 39 years.

[14] Subjects who failed the attention screen were different demographically from those who passed. Specifically, the group who failed included a significantly higher proportion of young and male respondents.

the number of units their workers produce. After defining the task, we introduce participants to their new employees, Worker A and Worker B, identifying the latter as the Target. The workers completed the same training course, and they operate independently in the factory. Couched in non-technical language about factory records, we explain that each worker's type, or true level of competence, is a random draw from a uniform distribution ranging from 950 to 1,450 units per week ($\mu_w \sim U[950, 1450] \; \forall \; w \in \{T, C\}$). Their weekly production follows a normal distribution with a fixed standard deviation and a mean equal to their type ($Y_{wt} \sim N[\mu_w, \sigma_w^2]$).[15] As the "game" begins, participants observe eight weeks of the Comparator's production before we introduce the Target worker for another eight weeks. Then they vote either to extend the Target's contract or to hire a replacement. If they vote to extend, the Target continues producing for another sixteen weeks. If they vote to replace, a new worker takes over production. We remind subjects that the replacement worker's type is also a draw from the uniform distribution described above. Having made their choice, subjects view a summary of factory output for weeks 17-32 and their corresponding cash bonus. For reference, Table 1 summarizes the flow of the performance voting task and notes specific differences across the signal extraction and signal recognition designs.

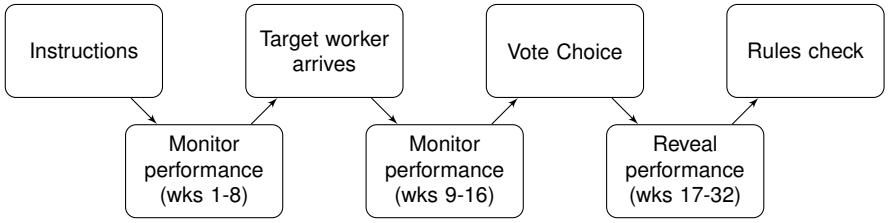
The defining element of the Signal Extraction design is that the participant never observes the Target's performance directly. The Comparator operates alone in the factory for eight weeks. The Target, delayed for reasons beyond his/her control, arrives in week nine, at which point the subject observes only their combined performance ($Y_{T,t} + Y_{C,t}$). This challenges participants to extract their Target's competence from a stream of performance data contaminated by a disturbance with a knowable distribution ($Y_C \sim N(\mu_C, \sigma_C)$). In Experiment 1 we randomize the variability of the Comparator's output — assigning subjects to either a high noise ($\sigma_{C,High} = 250$) or low noise ($\sigma_{C,Low} = 100$) arm — to assess the moderating effect of a low-noise disturbance on the ability to extract a competence signal. This closely and intentionally mimics the extraction challenge motivating the "competence" model of performance voting (Duch and Stevenson 2008; see also: Alesina and Rosenthal 1995).

The Signal Recognition task, by comparison, is a one-worker game that provides an unobstructed look at the Target's performance. Participants are asked to evaluate the performance of a

[15] Participants must click through each instructional screen; however, to encourage attention, they can only advance after a fixed interval, typically ten seconds. After reading the full instructions, participants can opt to go over them a second time. On the second reading, participants can click through as quickly (or slowly) as they like.

Table 1: *Overview of experimental designs*

A. Flow of the core task



B. Experimental interventions

Design	Randomizations	Performance weeks 1-8	Target Arrives	Performance weeks 9-16
Signal Extraction	• Worker types (μ_w)[a] • Comparator noise (σ_C)[b]	$Y_{C,t} \sim N(\mu_C, \sigma_C)$	T joins C	$(Y_{C,t} + Y_{T,t})$
Signal Recognition	• Worker types (μ_w)	$Y_{C,t} \sim N(\mu_C, 250)$	C departs, T replaces	$Y_{T,t}$

[a] Worker types are drawn from a uniform distribution: $\mu_w \sim U[950, 1450]\ \forall\ w \in \{T, C\}$.
[b] We assign subjects to the low variance Comparator ($\sigma_{C,low} = 100$) or high variance Comparator ($\sigma_{C,high} = 250$) condition.

single factory worker over the course of the sixteen week trial. However, we surprise participants after week eight with the news that their worker suddenly and unexpectedly left the factory. A new worker, now the Target, has already transferred from another department to replace the old worker, now the Comparator. The new aim is to vote after week sixteen to reappoint or replace the Target. In contrast to the Extraction design, there is no performance signal to extract. Subjects need only recognize the Target's performance over the eight weeks of observation as the relevant signal of competence. We encourage attention to this change by making the Comparator's departure an arresting event. The unexpected notice comes on a separate screen in bold letters with an explanation of how it changes the task. We also use a timer to prevent players from simply clicking though the notice. Both features are meant to break the cognitive focus on the Comparator's performance and signal the start of a new, more important stream of data. Finally, we make clear that the arriving Target worker is equally qualified for the job and should be judged by the same standards as the outgoing Comparator (i.e. $\mu_T \sim U[950, 1450]$ and $Y_T \sim N(\mu_T, \sigma_T)$).

Participants in our experiments, then, face two important challenges. First, and most importantly, they have to extract a signal about Worker B's competence from a stream of information

contaminated[16] by the Comparator's performance. Second, they have to apply their inference about the Target Worker to a decision with real financial consequence. The optimal decision rule, given the uniform distribution of worker type, is to extend the contract of any above-average worker ($\mu \geq 1200$) and replace any below-average worker ($\mu < 1200$). The instructions never define this or any decision rule, but participants largely understood the task. We used two questions to test comprehension of the uniform distribution of worker type and the normal distribution of weekly output. 81% answered one of the questions correctly and nearly half (48%) answered both correctly.[17]

Like Huber, Hill, and Lenz (2012), our experimental tasks depart from real-world politics in countless ways. Most significantly, to isolate the dynamics of performance evaluation, we sought to eliminate alternative judgmental dimensions that occupy voter attention in other contexts, particularly social identities (see especially Achen and Bartels 2016). At the same time, we mimic real-world electoral politics by presenting relevant performance information in the form of a stream, rather than a simple summary. Likewise, to enhance realism, the stream of information incorporates random variation in performance. Given our analytic purposes, our design's key analogy to the real world is the confounding (particularly in the Signal Extraction design) of the target worker's performance—whose type is unknown—by the performance of another. Importantly, while this exogenous influence is logically irrelevant to the evaluation in question, it reflects performance in the same domain by a comparable (indeed, in expectation, precisely equivalent) actor. In this way, we parallel the real-world confounding of the behavior of multiple governments in producing many important social outcomes.[18] Thus, while we cannot claim to have fully replicated the cir-

[16] The form of the "contamination" differs between Studies 1 and 2, as explained in subsequent sections.

[17] We asked subjects two questions at the conclusion of the task. The questions were:

> If a worker in this task averages 1,100 units per week, are they more likely to produce 1,300 or 1,500 units next week? (a) 1,300 units is more likely (correct); (b) 1,500 units is more likely; (c) 1,300 and 1,500 are equally likely; (d) Don't know
>
> If you replace a worker who averaged 1,200 units per week, is the new hire more likely to average 1,000 or 1,400 units per week? (a) 1,000 units is more likely; (b) 1,400 units is more likely; (c) 1,000 and 1,400 are equally likely (correct); (d) Don't know

The first question concerns the uniform distribution of worker type. 55% of subjects chose the correct answer. The second deals with the normal distribution of weekly payouts. 60% answered this question correctly. Our findings are robust to the inclusion or exclusion of those who did not answer the rules questions correctly.

[18] This feature of our design (i.e., the confounding of the performances of comparable actors) is an important contrast with Huber, Hill, and Lenz's (2012) examination of the impact of irrelevant information. In their experiment, the irrelevant information is, by design, presented separately, in a distinct format, and is explained as arising from a distinct process, specifically, the outcome of a "lottery". Further, the impact of the lottery outcome on the experimental participant's payout is an order of magnitude larger than the impact of the contribution of the evaluative target (i.e., the "allocator"). Finally, whereas participants are advised to ignore the irrelevant information in Huber, Hill,

cumstances of real-world performance voting, to the extent that our experiments capture crucial features of this important democratic task, they are a useful source of information about how voters judge performance in an interdependent world[19].

Baseline reactions to worker performance

Across our 1,937 respondents, 72% elected to extend their Incumbent's contract, and the baseline retention rate ranged from a low of 63% to a high of 76%. Figure 1 plots the proportion opting to extend the Incumbent as a function of each worker's performance across studies. The unconditional OLS estimates reveal, first, that respondents across all studies reacted to the Incumbent's performance. The upward slope of the estimates shows the proportion extending the Incumbent's contract increased with the Incumbent's underlying competence. This is true whether or not incumbent performance is obscured by the efforts of the Comparator–as in the extraction games. This is consistent with both the signal extraction and benchmarking hypotheses. However, the estimates in the right-hand panel suggest that respondents also gauge Incumbent competence in light on the efforts of the irrelevant Comparator. Specifically, subjects across studies appear to reward Incumbents – competent or otherwise – in the presence of a poor performing peer. This behavior is suggestive of a benchmarking response, especially in the signal recognition games where subjects observe the Incumbent's unobscured performance. Together, these results highlight a regular pattern of behavior across three studies and two game designs.

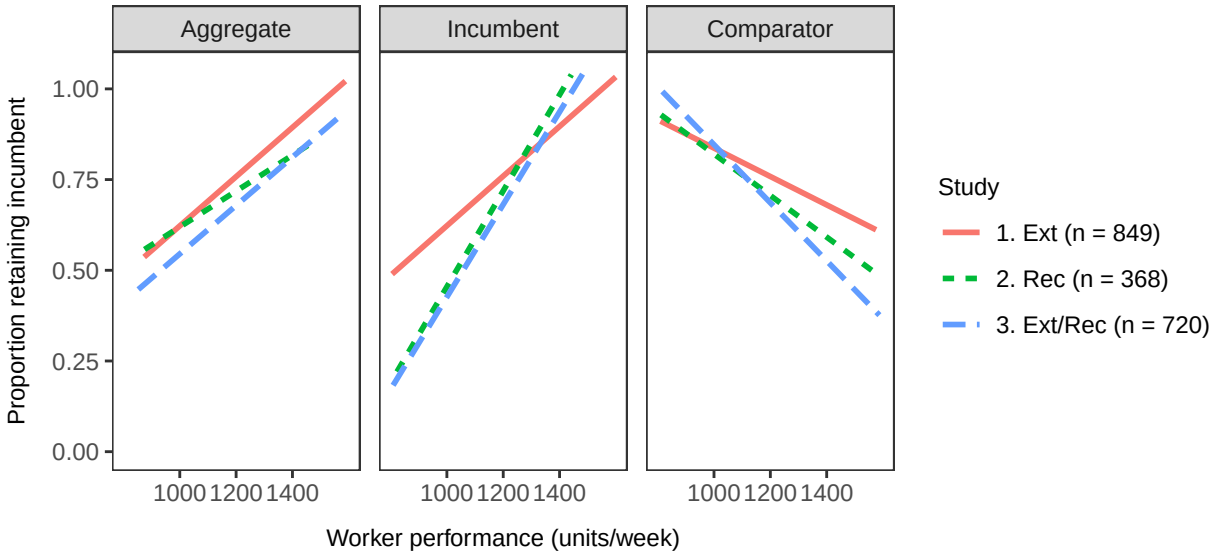
Experiment 1: Signal extraction

Experiment 1 represents our central test of the Signal Extraction hypothesis, including the moderating effect of disturbance variability on performance voting. The experiment uses the Signal Extraction design outlined in Table 1, and the game begins with one Comparator working alone in the factory for eight weeks. The Target worker arrives in week nine, at which point participants observe their workers' joint production ($Y_{T,t} + Y_{C,t}$) until they vote after week sixteen to either

and Lenz's (2012) experiment, our participants are *required* to attend to irrelevant information in order to extract information about the evaluative target's performance.

[19]Our reasoning in this regard parallels Lau and Redlawsk's (2006) justification of the "dynamic process-tracing methodology" for the study of the use of cognitive heuristics in election campaigns (pp. 63-64).

Figure 1: *Baseline response to worker performance*



Note: Lines are unconditional OLS estimates of the proportion of respondents opting to extend the incumbent's contract by each worker's competence (ie assigned type). Note that the length and mode of exposure to each worker's performance differs within and between study.

reappoint or replace the Target worker. In addition to randomizing each worker's underlying level of competence, we manipulate the variability of the Comparator's weekly output, assigning participants to either the "high-noise" treatment ($\sigma_{C,High} = 250$) or low-noise treatment ($\sigma_{C,Low} = 100$).[20] The variability of the Target's output is constant across treatment arms ($\sigma_T = 250$).

Given the design, we evaluate the signal extraction versus benchmarking theses in Experiment 1 by specifying the responsiveness to the Target and Comparator workers' competence, β_T and β_C, as follows:

$$y_i = \beta_T(\mu_{T,i}) + \beta_C(\mu_{C,i}) + \theta(\sigma_{C,Low}) + \alpha + u_i \tag{1}$$

where y_i is a binary indicator of subject i's vote to reappoint the Target; $\mu_{T,i}$ and $\mu_{C,i}$ represent the competence of the Target and Comparator workers respectively; $\sigma_{C,Low}$ indicates assignment to the low-noise condition; α is the intercept; and u_i is an error term. Note that μ_T and μ_C are strictly exogenous. Signal extracting behavior occurs when subjects vote on the Target's reappointment

[20] We selected the values of the variance parameters so that the task of extracting the Target worker's performance is neither trivially easy (especially in the low-noise condition) nor unreasonably difficult (especially in the high-noise condition). We also chose values that would prevent, in expectation, a sizable number of negative values of revealed performance.

Table 2: *Performance voting and the effect of a low-noise disturbance*

	DV: Extend the Target's Contract			
	(1) Baseline	(2) Response by treatment arm		
	Full sample	High noise	Low noise	Diff (Δ)
Target performance, β_T	0.076	0.057	0.092	
	(0.010)	(0.014)	(0.014)	
Low-noise effect, Δ_T				0.036
$P(\Delta_T \leq 0)$				0.033
Comparator performance, β_C	−0.032	−0.018	−0.043	
	(0.010)	(0.014)	(0.014)	
Low-noise effect, Δ_C				−0.025
$P(\Delta_C \geq 0)$				0.104
Low noise Comparator ($\sigma_C = 100$)	−0.019			
	(0.028)			
Intercept, α	0.241	0.301	0.152	
	(0.166)	(0.230)	(0.235)	
Hypothesis tests				
No performance vote $P(\beta_T \leq 0)$	0.000	0.000	0.000	
No benchmarking $P(\beta_C \geq 0)$	0.000	0.105	0.001	
Observations	849	849		
R^2	0.078	0.777		

Note: OLS estimates with standard errors in parentheses. Difference statistics, Δ, provide the basis for testing the moderating effect of a low-noise disturbance on the performance vote and benchmarking responses. We scale performance in 100s of units to facilitate interpretation.

according to his/her competence alone (ie $\beta_T > 0$ and $\beta_C = 0$). Benchmarking, on the other hand, is evident when individuals judge the Target's performance not in isolation but against the baseline of the Comparator (ie $\beta_T > 0$ and $\beta_C < 0$). What distinguishes these behaviors, then, is the response to the Comparator, β_C. Model 1 of Table 2 presents the estimates.

The results reveal that subjects reacted systematically to the Target worker's efforts. For every 100-unit increase in the Target's competence, the probability of reappointment increased by 0.076, and we can reject the null hypothesis of no performance voting ($P(\beta_T \leq 0) < 0.001$). This is consistent with the extraction thesis, and it is an important finding given that subjects must use their knowledge of the Comparator to extract information about the Target. It seems, then, that subjects "got the signal." However, participants also responded to the disturbance itself, shunning reappointment of the Target as the Comparator's competence increased ($\beta_C = -0.032$). Inconsistent with the extraction logic, we can also reject the null hypothesis of no benchmarking

($P(\beta_C \geq 0) < 0.001$). Together, the estimates reveal that participants, on average, are willing extend the contract of an incumbent — or Target worker — whose performance is poor in absolute terms so long as he or she outperformed the Comparator.

To what extent does the variability of the Comparator's output condition these reactions? To identify the moderating relationship, we specify the reappointment function for the low noise and high noise conditions:

$$y_i = \beta_{T,Low}(\mu_{T,i}) + \beta_{C,Low}(\mu_{C,i}) + \alpha_{Low} + u_i \tag{2}$$

$$y_i = \beta_{T,High}(\mu_{T,i}) + \beta_{C,High}(\mu_{C,i}) + \alpha_{High} + u_i \tag{3}$$

Then we test the increase in responsiveness to the Target and Comparator workers as:

$$\Delta_T = \beta_{T,Low} - \beta_{T,High} \tag{4}$$

$$\Delta_C = \beta_{C,Low} - \beta_{C,High} \tag{5}$$

If the variability of the disturbance moderates the performance vote, the core prediction of extraction theory, the response to the Target worker should be highest among subjects assigned to the low-noise arm ($\Delta_T > 0$). In contrast, benchmarking behavior is evident when the low-noise disturbance facilitates a clear comparison, implying greater responsiveness to *both* workers ($\Delta_T > 0$ and $\Delta_C < 0$). To evaluate the moderating effect of the low-noise treatment, we estimate the parameters in Equations 2 and 3 using OLS.[21] We conduct post-hoc tests of significance for the differences identified in Equations 4 and 5. Model 2 of Table 2 presents the estimates.

We find that clear disturbances do amplify the response to incumbent performance. Specifically, the effect of the Target, or incumbent, worker's competence is significantly higher among subjects assigned to the low noise condition than among those in the high noise condition ($\Delta_T = 0.036$, $P(\leq 0) = 0.033$). Although this is consistent with central prediction of the extraction thesis, we also observe greater sensitivity to the Comparator's competence in the low-noise condition ($\Delta_C = -0.025$, $P(\geq 0) = 0.104$). Consistent with a benchmarking response, we see that respondents are either more willing or more able to gauge the Target's performance against their Comparator's when

[21] We do so in a single regression by including an indicator variables for treatment arm and interacting those with the performance of the Target and Comparator

the Comparator's efforts are clearest. This is the opposite of what we would expect if the extraction thesis reflected the process by which individuals cope with the problem of interdependence.

Experiment 1 participants responded to both the signal and the noise, and the cost is notable. Subjects evaluating the Target purely at random are expected to earn a bonus of \$0.672.[22] Perfect application of the optimal decision rule (replacing a Target if $\mu_T < 1,200$) should net a bonus of about \$0.69. Experiment 1 participants earned an average of just \$0.674. This is not a sign of innumeracy, inattentiveness, or even suboptimal application of the decision rule. Rather, the failure to discard irrelevant information depressed bonuses to the level expected had participants chosen at random ($P(\mu_{bonus} \geq 0.672) = 0.177$).

We designed the Experiment 1 task as a direct test of signal extraction theory. Although we find that individuals capably parse streams of performance data into endogenous and exogenous components, there is, at best, incomplete support for extraction as a model of evaluative judgment in the face of interdependence. Individuals systematically fail to discard the irrelevant information they glean about Comparators. Yet this is the core of the extraction thesis. What we find instead is that individuals use the exogenous information as a baseline against which to evaluate their incumbent, or Target worker, and this comparative response intensifies when the performance signal is clearest. However, we worry that the experimental task may unintentionally promote comparison between the workers. After all, the task of extracting the Target worker's type in the Signal Extraction design is by definition an exercise in differentiation. We designed the Signal Recognition task with this analytical consideration in mind.

Experiment 2: Signal recognition

Experiment 2 follows the Signal Recognition design described previously. Here we ask participants to evaluate the performance of a single worker over a sixteen week trial period. However, the worker unexpectedly leaves the factory after week eight and is replaced by a transfer from another department. The departed worker, then, becomes the Comparator and the incoming transfer the Target, and subjects must elect after week sixteen to reappoint or to replace the Target worker. For

[22] Given the uniform distribution of worker types, a worker selected at random ought to average 1,200 units per week ($E[Y_w] = 1,200 \ \forall \ w$). Therefore, the expected total, T, for two workers over 32 weeks (minus the Target's eight-week delay) is $E[T] = 32 * E[Y_C] + 24 * E[Y_T] = 56 * 1,200 = 67,200$. At \$1/100,000 units, the expected bonus given random selection is \$0.672.

Table 3: *Response to unobstructed performance and disturbance*

	DV: Extend the Target's contract	
	Observed output[a]	Worker types[b]
Target performance, β_T	0.123	0.123
	(0.013)	(0.015)
Comparator performance, β_C	-0.055	-0.058
	(0.012)	(0.015)
Intercept	-0.105	-0.064
	(0.214)	(0.264))
Hypothesis tests		
No performance vote, $P(\beta_T \leq 0)$	0.000	0.000
No benchmarking, $P(\beta_C \geq 0)$	0.000	0.000
Observations	368	368
R^2	0.233	0.192

Note: OLS estimates with standard errors in parentheses. We scale performance measures in 100s of units to facilitate interpretation.
[a] We measure relevant performance as B's average output in weeks 9 to 16 and the disturbance as A's average output in weeks 1 to 8.
[b] We use each worker's assigned type to measure relevant performance and disturbance.

reference, each workers underlying competence, or type is random a draw from a uniform distribution ($\mu_w \sim U[950, 1450]$), and weekly output follows a normal distribution with a mean equal to their competence and a fixed standard deviation ($Y_w \sim N[\mu_w, 250]$).

This one-worker design is meant to discourage comparative evaluation and instead promote a "clean" evaluation of the Target worker's performance. In stark contrast to the Extraction task, there is no signal to extract in Experiment 2, obviating entirely the need for inference. Participants observe the Target in isolation, and weekly output, though variable *is* the uncontaminated competence stream. If, in this context, individuals still rely on the Comparator's performance as a baseline for their reappointment decision, it will be strong evidence in support of a benchmarking response and against signal extraction. Table 3 presents OLS estimates of Equation 1 for the 368 participants in Experiment 2.

As with Experiment 1, the results are consistent with the benchmarking hypothesis. On the one hand, we find that the probability of reappointment increases in direct proportion to the Target's actual level of competence ($P(\beta_T \leq 0) < 0.001$). This performance voting response is consistent with both benchmarking and extraction behavior. However, the Comparator's efforts enter the vote decision as a baseline for comparison rather than an exogenous disturbance. It is unlikely we

observe this response by chance alone ($P(\beta_C \geq 0) < 0.001$). Together, the estimates reveal that respondents successfully identified the Target's performance but judged it in part for its distinction from a Comparator. This means that a below average Target is rewarded for outpacing an even less competent Comparator.

If the results in Experiment 1 were merely a function of the task design, the benchmark response should attenuate in Experiment 2, which is designed explicitly to inhibit benchmarking. Yet, the response is larger and more precisely estimated. This is consistent with the finding from Experiment 1 of greater benchmarking in the presence of a low-noise disturbance. In both instances, simpler evaluative tasks prompted more—not less—reliance on a reference point. This should not be the case if the comparative response is just a crutch of convenience in a contrived information environment. This suggests that, in the context of interdependent performance, citizens rely on available benchmarks, whether they are relevant or not. The observed reduction in real-world performance voting in more informationally challenging circumstances, then, may reflect the simple and bias-prone challenge of identifying a benchmark, rather than a sophisticated response to the challenge of signal extraction.

Experiment 3: Seeking a signal

Do citizens *want* to benchmark government performance? We find clear evidence of a benchmarking response in Experiments 1 and 2 – subjects judged their incumbent or Target worker's performance not in absolute terms but in contrast to a Comparator's performance. Whether or not individuals intuitively and/or actively pursue this type of benchmark comparison is the focus of Experiment 3. We address this question by allowing participants to uncover summary reports of their workers' performance. Given this opportunity, do participants seek relevant information about the Target worker, or do they gravitate toward comparative/benchmarked indicators of performance?

Experiment 3 assigned participants at random to complete either a Signal Extraction task or a Signal Recognition task with probabilities $\frac{2}{3}$ and $\frac{1}{3}$ respectively.[23] The key difference in Experiment 3 is that participants have the chance to view summary information about their workers' production.

[23] See the earlier notes and Table 1 for details on study designs.

Immediately prior to voting to reappoint or replace the Target worker, subjects have the opportunity to view as many as two of the following reports, presented in random order:

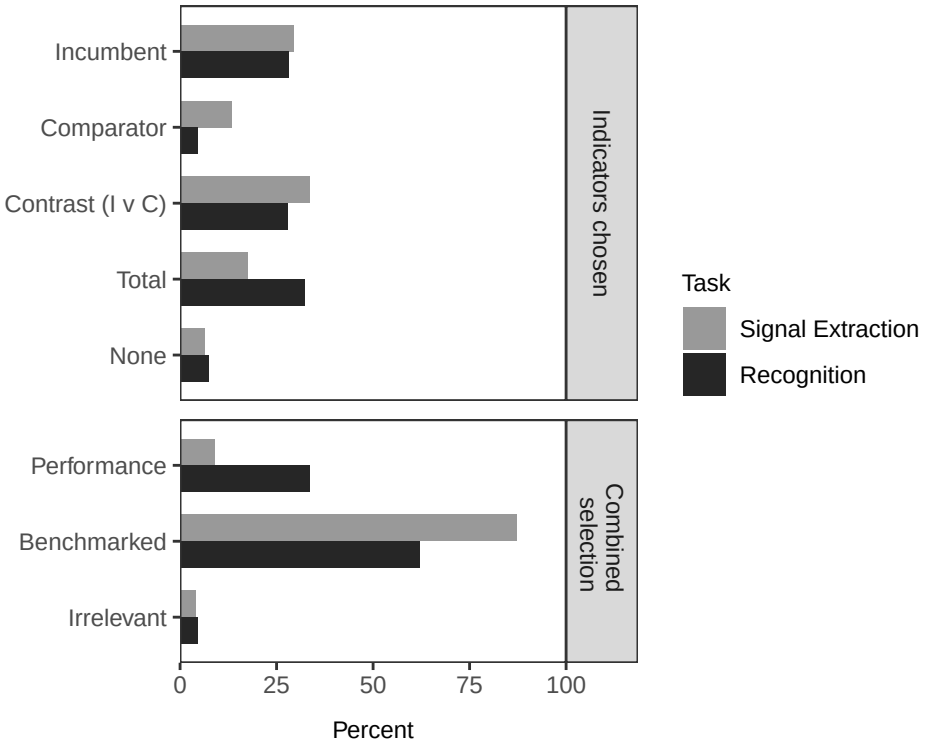
1. Target worker's average production [*Target report*]
2. Comparator's average production [*Comparator report*]
3. Comparison of the Target and Comparator's productivity [*Contrast*]
4. Total units produced and bonus to date. [*Yield*]

Only the first report—the Target's average—contains relevant performance data. In fact, it eliminates the need for evaluative judgment altogether. The other reports are irrelevant (again, the Comparator's performance is strictly exogenous). That said, reports two and three should appeal to individuals looking to judge performance by way of comparison.

Note that our goal in Experiment 3 is to *observe* information seeking behavior. Given this option, do individuals pursue benchmarked performance data even if the available comparison is irrelevant? If the benchmark effect we observe in the prior studies is an unfortunate response to the task design, Experiment 3 participants ought to eschew comparative information and seek instead the direct report of the Target's performance. This should also mitigate the magnitude of the benchmark response. However, if benchmarking is a general mode of decision making, access to comparative information should reveal this tendency.

Figure 2 plots the distribution of reports chosen by study design. Because participants had the option to select up to two reports, the top panel presents the selections in isolation and the bottom characterizes the joint distribution. The results reveal a clear preference for comparative, or benchmarked, performance summaries. Beginning in the top panel, the Contrast report is as or more frequently chosen than the Target's report and is the modal choice among subjects assigned to the Extraction task (34%). Turning to the joint selection of reports in the bottom panel, a majority of respondents in the Recognition arm (63%) opted for some form of comparison, either viewing the Target and Comparator reports separately or the Contrast report. In the Extraction condition, which required participants to extract information about the Target's performance, nearly every participant (87%) sought a comparison. Less than 10% of respondents in the Extraction arm and one-third (33%) in the Recognition arm chose the Target's report without also selecting the Contrast or Comparator's report. Critically, only one in twenty-five choose strictly irrelevant reports (neither

Figure 2: *Expressing a preference for benchmarking*



Note: The top panel shows the relative frequency of individual reports chosen. The bottom panel characterizes the specific combination of reports chosen: performance only (Target report absent any comparison), benchmarked (any combination of Target and Comparator performance), and irrelevant (chose no information about the Target worker).

the Target's average nor the comparison).[24] These results support the idea that benchmarking is an ingrained mode of performance evaluation.

Though the preference for a benchmark is clear, the pattern of information seeking differs by task. Those assigned to the Study 3 Extraction arm were more likely to select comparative reports. The difference in the distributions across task designs is statistically significant ($\chi^2_2 = 70.3$, $p < 0.001$). We suspect that this is the result of the relative simplicity of the Recognition task: view the Target's performance in isolation and then judge the Target. It stands to reason that the reports are less valuable in this context. This may explain, for instance, why the Yield report was the modal choice in the Recognition arm: participants may only indulge their impatience to know this information—which is strictly irrelevant to the task—if they are confident they already have the performance information they need. In contrast, participants assigned to the Extraction

[24] This is far lower than the 36% we would expect if participants were selecting reports at random.

task confront a genuine signal extraction problem: infer their Target worker's competence from a stream of data obscured by an exogenous disturbance. The necessity of judging performance amid uncertainty elevates the utility of the summary reports. In this context, nearly every respondent sought comparative information.

One might also wonder if the pattern of responses reflects an undue pressure to select two reports. If participants saw the choice of *up to two* reports as a mandate rather than an option, the results might inflate the percentage "seeking" comparative information and understate the percentage who would be happy viewing only the Target worker's average production. If this were true, respondents who chose to view the Target's report should select their other report at random. Yet the choice is systematic ($\chi^2_3 = 156.1$, $p < 0.001$ and $\chi^2_3 = 58.5$, $p < 0.001$ for the Extraction and Recognition conditions respectively).[25] Among respondents in the Extraction task, for instance, 85% of those who chose to view the Target's average also selected a comparative report.

Conclusion

Confronted with a world in which important outcomes reflect interdependent processes beyond the jurisdiction of any one government, the performance-oriented voter must adopt a strategy for inferring incumbent competence – for identifying the unique contribution to inherently composite outcomes of actors whose performance the voter seeks to optimize. While some research indicates voters can successfully separate endogenous performance from exogenous influences – i.e., that voters can *extract* an incumbent competence signal from performance information contaminated with the effects of irrelevant variables – other research suggests voters cognitively economize by *benchmarking* their incumbent's performance relative to that of comparable actors in other jurisdictions. Still other scholars are skeptical that voters can judge performance effectively in any domain, whether it involves interdependent outcomes or not, and suspect performance voting often amounts to no more than *blind retrospection*.

In our studies, in which participants engaged in performance voting in a setting approximating the real-world analog in critical features, we find a "voter" whose behavior amalgamates aspects of the three strategies suggested by the literature. Our participants can extract a relevant performance

[25] The same is true if we examine the distribution over the remaining reports, excluding the option of refusing a second report.

signal from a noisy stream of information, but seem unable to set aside the noise and focus solely on the signal. Rather, they consistently engage in performance comparisons that are, given the design of our studies, logically irrelevant to the performance evaluation with which participants were tasked. This benchmarking behaviour emerges whether extracting performance information is hard or easy (Study 1), and even when no extraction of information is necessary and attention to comparators is actively discouraged (Study 2). The opportunity to acquire summary information about performance – which obviates the need for evaluative inference of any kind – also does not eliminate evidence of benchmarking (Study 3). On the contrary, the opportunity to acquire such information often promotes (and never reduces) benchmarking. Participants also actively choose summary information that facilitates comparison, even those participants who otherwise acquire information that is completely diagnostic of performance.

The implications of these findings for the problem of unmasking accountability for interdependent outcomes, and for the study of performance voting generally, are complex. That participants in the Extraction task appear able to identify an actor's performance from a noisy stream of information is remarkable, even given the relative simplicity of the task (compared to real-world performance voting) and the presence of clear incentives for high-quality judgement. In an otherwise similar setting, past research suggests people have difficulty separating a stream of performance outcomes in the presence of information about a single, clearly distinct and obviously irrelevant outcome (Huber, Hill, and Lenz 2012, pp. 731-734). Our participants, in contrast, seemingly extract information about the performance of not one, but two actors. Having parsed the relevant variables, however, our "performance voters" fail to focus their judgemental attention where it belongs: on the performance of the actor whose competence is under evaluation.

To be sure, comparing the performance of a given actor to a benchmark is perfectly reasonable under the right conditions, specifically, if the benchmark captures the average (or other relevant parameter) of a relevant distribution of performance. This may be the case for the voter who compares national economic performance to a sensible aggregation of global comparators, as Kayser and Peress (2012) suggest.[26] In our studies, however, benchmarking occurs with respect to a

[26]Note, however, that the normative attractiveness of this approach depends on the relationship between the *observed* distribution of performance and some normatively acceptable performance standard. That is, if all observed comparators underperform, by some criterion, then the fact that a given incumbent is above average does not necessarily justify their retention.

comparator who is explicitly irrelevant to the judgement at hand. Worse, the comparison occurs despite the fact participants are informed of the normatively relevant benchmark – the distribution of worker types – at the start of each task and then reminded of this information just prior to "voting" on worker retention. Consistent with Study 3's results on the propensity to actively seek comparative information, the participants in our studies seemingly have a "taste" for benchmarking, even when it is manifestly unnecessary and, indeed, counterproductive.

To the extent this benchmarking pattern generalizes to the real world, these results have important implications. Building on Kayser and Peress (2012), our findings help to reconcile the deep contradiction between robust evidence consistent with ultra-rational signal extraction behavior (e.g. Duch and Stevenson 2008), on the one hand, and findings consistent with ultra-irrational blind retrospection behavior (e.g. Achen and Bartels 2016), on the other. Benchmarking explains how a world of voters engaged in cognitively cheap performance evaluation can nevertheless generate cross-national differences in performance voting that approximate the expectations of a much more cognitively demanding judgemental model. In partial contrast to Kayser and Peress (2012), we show that benchmarking does not depend on, what they call, "pre-benchmarked" information (p. 662). Instead, our evidence suggests voters may benchmark performance quite automatically, provided a salient benchmark is available. Like any useful heuristic, however, benchmarking can also be a source of biased judgements (Tversky and Kahneman 1981), as it is in our studies. Put simply, bad benchmarks will lead to bad judgements.

For those interested in how voters evaluate performance in the context of interdependence, the conditionality of "quality benchmarking" on the availability of "quality benchmarks" suggests a host of new research questions. What benchmarks are salient in the domains of performance important to vote choice? Is there systematic variation in the quality of these benchmarks across domains? Economic voters seem doubly blessed in this regard: suitable benchmarks (e.g., global and regional GDP growth statistics) are regularly reported, and evaluative standards in the economic domain are fairly consensual. In other areas, good performance may be both less visible and harder to assess. Further, within a given domain of performance, how stable are benchmarks over time, and how variable across space? Kayser and Peress's (2012) estimates of benchmarking of national economic performance, which cover as many as 22 countries and more than 60 years, are consistent with a high degree of stability and cross-national similarity of benchmarks in this domain. On other

dimensions of performance, however, a paucity of information and ambiguous standards may imply greater heterogeneity and the potential for political manipulation.

References

Achen, Christopher H., and Larry M. Bartels. 2016. *Deomocracy for Realists: Why Elections do not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Alcañiz, Isabella, and Timothy Hellwig. 2011. "Who's to Blame? The Distribution of Responsibility in Developing Democracies". *British Journal of Political Science* 41 (2): 389–411.

Alesina, Alberto, and Howard Rosenthal. 1995. *Partisan Politics, Divided Government, and the Economy*. Cambridge: Cambridge 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 (1): 271–296.

Arel-Bundock, Vincent, Andre Blais, and Ruth Dassonneville. 2018. "Do Voters Benchmark Economic Performance?" *British Journal of Political Science*: 1–13.

Aytac, Selim Erdem. 2017. "Relative Economic Performance and the Incumbent Vote: A Reference Point Theory". *Journal of Politics* 80 (1): 16–29.

Besley, Timothy, and Anne Case. 1995. "Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition". *American Economic Review* 85 (March): 25–45.

Bodet, Marc-André, Melanee Thomas, and Charles Tessier. 2016. "Come Hell or High Water: An Investigation of the Effects of a Natural Disaster on a Local Election". *Electoral Studies* 43 (): 85–94.

Campello, Daniela, and Cesar Zucco Jr. 2016. "Presidential Success and the World Economy". *The Journal of Politics* 78 (2): 589–602.

Carlin, Ryan E., Gregory J. Love, and Cecilia Martínez-Gallardo. 2015. "Security, Clarity of Responsibility, and Presidential Approval". *Comparative Political Studies* 48 (4): 438–463.

Case, Anne, James R. Hines, and Harvey S. Rosen. 1993. "Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States". *Journal of Public Economics* 52 (3): 285–307.

Casey, Logan S., et al. 2017. "Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection". *SAGE Open* 7 (2): https://doi.org/10.1177/215824401

Cole, Shawn, Andrew Healy, and Eric Werker. 2012. "Do voters demand responsive governments? Evidence from Indian disaster relief". *Journal of Development Economics* 97 (2): 167 –181.

Cook, Philip J., and Anthony A. Braga. 2001. "Comprehensive Firearms Tracing: Strategic and Investigative Uses of New Data on Firearms Markets". *Arizona Law Review* 43:277–309.

Duch, Raymond M., and Randolph Stevenson. 2008. *The Economic Vote: How Political and Economic Institutions Condition Election Results*. New York: Cambridge University Press.

— . 2010. "The Global Economy, Competency, 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.

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes". *The Quarterly Journal of Economics* 123 (2): 703–745.

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

Fowler, Anthony, and Andrew B. Hall. 2018. "Do Shark Attacks Influence Presidential Elections? Reassessing a Prominent Finding on Voter Competence". *The Journal of Politics* 80 (4): 1423–1437.

Fowler, Anthony, and B. Pablo Montagnes. 2015. "College football, elections, and false-positive results in observational research". *Proceedings of the National Academy of Sciences* 112 (45): 13800–13804.

Gelineau, Francois, and Karen Remmer. 2006. "Political decentralization and electoral accountability: The Argentine Experience, 1983–2001". *British Journal of Political Science* 36 (1): 133–157.

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

Hart, Austin. 2016. *Economic Voting: A Campaign-Centered Theory*. New York: Cambridge University Press.

Hayes, Rosa C., Masami Imai, and Cameron A. Shelton. 2015. "Attribution error in economic voting: Evidence from trade shocks". *Economic Inquiry* 53 (1): 258–275.

Healy, Andrew, and Gabriel S. 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. 2010. "Random Events, Economic Losses, and Retrospective Voting: Implications for Democratic Competence". *Quarterly Journal of Political Science* 5:193–208.

Healy, Andrew, and Neil Malhotra. 2013. "Retrospective Voting Reconsidered". *Annual Review of Political Science* 16 (1): 285–306.

Healy, Andrew, Neil Malhotra, and Cecilia Hyunjung Mo. 2010. "Irrelevant Events Affect Voters's Evaluations of Government Performance". *Proceedings of the National Academy of Sciences* 107 (29): 12804–12809.

Hellwig, Timothy. 2008. "Globalization, Policy Constraints, and Vote Choice". *Journal of Politics* 70 (4): 1128–1141.

— . 2015. *Globalization and Mass Politics: Retaining the Room to Maneuver*. New York: Cambridge University Press.

Hobolt, Sara, James Tilly, and Susan Banducci. 2013. "Clarity of responsibility: How government cohesion conditions performance voting". *European Journal of Political Research* 52:164–187.

Holmstrom, Bengt R. 1982. "Moral Hazard in Teams". *Bell Journal of Economics and Management Science* 13 (2): 324–340.

Huber, Gregory A., Seth J. Hill, and Gabriel S. Lenz. 2012. "Sources of Bias in Retrospective Decision Making: Experimental Evidence on Voters's Limitations in Controlling Incumbents". *American Political Science Review* 106 (4): 720–741.

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

— . 2019. "Benchmarking across Borders: An Update and Response". *British Journal of Political Science*: 1–4.

Kiewiet, D. Roderick. 1983. *Macroeconomics and Micropolitics: The Electoral Effects of Economic Issues*. Chicago: University of Chicago Press.

Kinder, Donald R., and E. Roderick Kiewiet. 1981. "Sociotropic Politics: The American Case". *British Journal of Political Science* 11 (2): 129–161.

Kramer, Gerald H. 1971. "Short-Term Fluctuations in U.S. Voting Behavior, 1896-1964". *American Political Science Review* 77 (1): 92–111.

Ladd, Jonathan. 2007. "Predispositions and Public Support for the President during the War on Terrorism". *Public Opinion Quarterly* 71 (4): 511–538.

Larsen, Martin. 2018. "Is the Relationship Between Political Responsibility and Electoral Accountability Causal, Adaptive and Policy-Specific?" *Political Behavior* (). doi:10.1007/s11109-018-9483-3.

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. New York: Cambridge University Press.

Leigh, Andrew, and Mark McLeish. 2009. "Are State Elections Affected by the National Economy? Evidence from Australia". *Economic Record* 85:210 –222.

Lenz, Gabriel S. 2012. *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.

León, Sandra. 2012. "How do citizens attribute responsibility in multilevel states? Learning, biases and asymmetric federalism. Evidence from Spain". *Electoral Studies* 31 (1): 120–130.

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

Norpoth, Helmut. 2001. "Divided Government and Economic Voting". *Journal of Politics* 63 (2): 414–435.

Pierce, Glenn L., et al. 2004. "The Characteristics and Dynamics of Illegal Firearms Markets: Implications for a Supply-Side Enforcement Strategy". *Justice Quarterly* 21 (2): 391–422.

Powell, G. Bingham. 2000. *Elections as instruments of democracy: Majoritarian and proportional visions*. New Haven: Yale University Press.

Powell, G. Bingham, 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.

Samuels, David. 2004. "Presidentialism and Accountability for the Economy in Comparative Perspective". *American Political Science Review* 98 (3): 425–436.

Shipan, Charles R., and Craig Volden. 2008. "The Mechanisms of Policy Diffusion". *American Journal of Political Science* 52 (4): 840–857.

Simonovits, Gabor. 2015. "An Experimental Approach to Economic Voting". *Political Behavior* 37 (4): 977–994.

Tavits, Margit. 2007. "Clarity of Responsibility and Corruption". *American Journal of Political Science* 51 (1): 218–229.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures". *Journal of Political Economy* 64 (5): 416–424.

Tversky, Amos, and Daniel Kahneman. 1981. "Judgment under Uncertainty: Heuristics and Biases". *Science* 185 (4157): 1124–1131.

Wolfers, Justin J. 2007. "Are Voters Rational? Evidence from Gubernatorial Elections". Working paper. Available at: https://users.nber.org/ jwolfers/papers.

Supplemental materials for "Unmasking Accountability"

Austin Hart* Scott Matthews †

November 6, 2019

Summary

*School of International Service, American University. Email: ahart@american.edu
†Department of Political Science, Memorial University. Email: scott.matthews@mun.ca

1 Recruitment and sample demographics

1.1 Recruitment on MTurk

We recruited subjects for this study using Amazon.com's Mechanical Turk (AMT) marketplace. We offered AMT workers over the age of 18 and operating with a U.S. IP address $0.30 to play a decision making "game" that would take 5-7 minutes to complete. Specifically, those who clicked on our request saw the prompt below:

> **You may only participate once in this study.**
>
> We are conducting an academic study about how people make decisions. We're asking you to play a short game in which you'll evaluate the performance of hypothetical factory workers. The game will take 5-7 minutes to play. In addition to the base payment for this HIT, you will also earn a bonus based on the quality of your game play (average bonus = $0.67).
>
> Select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.
>
> Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

To reduce bias in the type of worker responding to the request, we automated calls for 50-100 respondents at fixed intervals over many days. We limited calls to 10am - 10pm EST. We used an IP filter in Qualtrics and an embedded Java screen in AMT to prevent repeated participation. We limit the sample to AMT workers with 95% approval ratings or better.

1.2 Sample characteristics and attentiveness

From August 9 to September 13, 2018, 2,452 AMT workers ($n_1 = 1,001$, $n_2 = 511$, $n_3 = 940$) followed the link to the task, hosted in Qualtrics. 99% consented to participate. We removed about 18% of respondents for failing a screener question designed to limit "click-through" behavior. The screener asked respondents for a non-obvious answer to a simple question:

> We would like to improve the design of our questionnaires. To demonstrate that you've read this much, please select Opera from the list below. Yes, ignore the question and select this option.
>
> Which web browser are you using to complete this study?
>
> - Chrome
> - Firefox
> - Explorer
> - Safari
> - Opera
> - Other/Not sure

In total, 1,937 respondents completed one of the three studies. The median time to completion was 5 minutes, 14 seconds. Table 1 describes the characteristics of each sample based on self-responses to four demographic questions. Overall, participants were 50% female and 78% white with a median age between 30 and 39. Chi-squared tests of independence reveal significant differences in Education across samples. However we do not anticipate that these differences influence the mode

of information processing.

Table 1: *Demographic characteristics by study*

		Relative Frequency (%)				Difference	
		Study 1	Study 2	Study 3			
Item	Response	Extract	Recog	Extract	Recog	$\chi^2(df)$	$Pr(>\chi^2)$
Age	18-29	33.9	39.4	34.5	32.1	14.2(12)	0.29
	30-39	36.2	34.5	37.0	41.7		
	40-49	15.9	13.9	13.5	12.2		
	50-65	12.5	10.1	11.6	11.8		
	+65	1.5	2.2	3.4	2.0		
Educ	No High School	0.5	0.8	0.8	0.8	37.8(18)	0.004
	High school/GED	10.4	9.2	9.5	10.5		
	Some college	26.1	19.3	22.9	23.5		
	2 yr degree	14.5	11.4	11.4	10.5		
	4 yr degree	37.1	44.3	36.0	35.6		
	Prof degree	10.0	10.6	15.7	16.2		
	Doctorate	1.4	4.3	3.6	2.8		
Race	Asian/Pac Islander	7.2	7.1	7.0	5.7	23.3(15)	0.08
	Black/Af American	6.5	6.8	7.4	8.5		
	Hispanic/Latino	5.0	7.9	5.3	6.5		
	Native Am/Am Indian	0.2	2.2	0.4	1.2		
	White, non-Hispanic	79.8	74.4	79.1	77.2		
	Other	1.3	1.6	0.6	0.8		
Sex	Female	51.2	46.3	49.9	51.0	2.6(3)	0.46
	Male/Other	48.8	53.7	50.1	49.0		

Note: Responses from 1,937 participants in total.

1.3 Payment

In addition to the $0.30 base payment, participants earned bonus payments proportional to their workers' output over sixteen weeks. Given the uniform distribution of worker type and normal distribution of weekly output, the mathematical expectation (i.e. the average payout for participants evaluating worker B at random) is $0.67 in the Extraction task. Individuals applying the optimal decision rule without error ought to exceed $0.68. By contrast, participants in the Study 1 and Study 3 Extraction tasks averaged just $0.66.

We observed a similar pattern in Study 2. The mathematical expectation for the bonus payment is $0.64 and the optimal application of the decision rule should raise this to $0.66. Participants in the Study 2 and Study 3 Recognition tasks averaged a payout of $0.64.

2 Instructions & comprehension

Respondents in both studies were exposed to the instructions over multiple screens. Subject to a 10-second timer, individuals manually clicked through each screen. Given the opportunity to read the rules again (with no timers) or to proceed to the task, 2% chose to review the rules. Here we present the text of the rules in each task, the alert about Worker A's departure in the Signal Recognition task, and the prompt to view performance reports in Study 3. We then present evidence that respondents generally understood the rules.

2.1 Rules, Signal Extraction Task

Rules of the Game

In this game you will serve as a factory supervisor in charge of evaluating employee performance.

You will learn about two workers and observe their performance over 16 weeks. Based on how many "units" they produce for the factory, you will then decide whether to extend one of your worker's contracts for another 16 weeks or hire a replacement.

Your goal is to choose workers in order to produce the most units. At the end of the game, you will earn a $1 bonus for every 100,000 units your workers produce.

About the Workers (1 of 3)

You hired two new employees, **A** and **B**. All new hires must complete a training course to prepare them to work in the factory. Both **A** and **B** completed the same course at the same school.

While it is not yet clear how A and B will perform on the job, your factory has employed many graduates from their school.

Records show that employees from their school averaged between 950 and 1,450 units per week. The distribution of average output is uniform across that range, meaning that it is equally likely that a graduate's weekly average is 950 units, 1,450 units, or any number in between.

About the Workers (2 of 3)

The new workers, **A** and **B**, work independently. They operate different machines, and the performance of one worker is unrelated to the performance of the other.

Factory records show that a worker's output varies each week for reasons beyond their control. Whether the worker is above average or below average, weekly production tends to follow a normal (bell-shaped) distribution. This means that, in a given week, a worker who averages 1,000 units is more likely to produce 900 or 1,100 units than to produce 700 or 1,300 units.

Because everyone completes a training course before starting at the factory, a worker's average output and week-to-week pattern of production typically does not change over time.

About the Workers (3 of 3)

Note that worker **A** will begin immediately. For reasons beyond their control, worker **B** cannot begin work until week 9.

After 16 weeks, you will have to decide whether to extend worker **B**'s contract for another 16 weeks or hire a replacement for this worker.

2.2 Rules, Signal Recognition Task

Rules of the Game

In this game you will serve as a factory supervisor in charge of evaluating employee performance in your department.

You will learn about a new worker and observe their performance over 16 weeks. Based on how many "units" they produce, you will then decide whether to extend the worker's contract for another 16 weeks or hire a replacement.

Your goal is to choose workers who produce the most units. At the end of the game, you will earn a $1 bonus for every 60,000 units your worker produces.

About the Workers (1 of 2)

You just transferred an employee, worker **A**, from another department. Like all employees in the factory, worker **A** is fully trained to work in any department.

While it is not yet clear how **A** will perform on the job, factory records show that employees transferred to your department average between 950 and 1,450 units per week. The distribution of average output is uniform across that range, meaning that it is equally likely that a worker's weekly average is 950 units, 1,450 units, or any number in between.

About the Workers (2 of 2)

Factory records show that a worker's output varies each week for reasons beyond their control. Whether the worker is above average or below average, weekly production tends to follow a normal (bell-shaped) distribution. This means that, in a given week, a worker who averages 1,000 units is more likely to produce 900 or 1,100 units than to produce 700 or 1,300 units.

Because everyone completes a training course before starting at the factory, a worker's average output and week-to-week pattern of production typically does not change over time.

After 16 weeks, you will have to decide whether to extend worker **A**'s contract for another 16 weeks or hire a replacement for this worker.

Note: screen presented in period 9

****ATTENTION: For reasons beyond their control, worker A has left the factory.****

Though unexpected, this should not affect production. The factory owner has already transferred in Worker **B** as a replacement. While it is not known how **B** will perform, remember that your records show that employees average between 950 and 1,450 units per week.

After week 16, you will now have the chance to renew **B**'s contract or to bring in a new worker.

2.3 Signal Search, Study 3

Note that Study 3 assigns participants at random to play either the Extraction or Recognition task. As such, the rules in each condition are presented as above. The one difference in the Study 3 version is the option to view performance reports. The question, shown below, appears immediately prior to the vote decision after week 16.

Note: screen presented prior to vote choice

In a moment, you will decide whether to extend Worker **B**'s contract or hire a replacement. Before you make your choice you may wish to review some performance information.

You may select up to two (2) reports.

- A's average production
- B's average production
- Total units produced and bonus earned to date
- Comparison of A and B's productivity

2.4 Comprehension

After the vote we posed two questions to gauge participants' understanding of the rules:

- If you replace a worker who averaged 1,200 units per week, is the new hire more likely to average 1,000 or 1,400 units per week?
- If a worker averages 1,100 units per week, are they more likely to produce 1,300 or 1,500 units next week?

The first concerns the uniform distribution of a worker's average production, and the second focuses on the normal distribution of weekly production. Table 2 presents the wording of each question and the distribution of responses by study. In total, 84% of respondents answered at least one question correctly, and 36% answered both questions correctly. This distribution for the full sample is significantly different than what we would expect if participants were answering the rules questions at random ($\chi^2 = 286(2), p < 0.001$). However, there are differences in comprehension on both items across studies. The higher percentage of correct responses in the Extraction design, especially Study 1, may reflect its difficulty relative to the Signal Recognition task. It is important to note here that rules comprehension has an inconsistent effect on the response to treatment and, to the extent that it does, comprehension magnifies the benchmark effect. See the section on robustness checks for model estimates.

Table 2: *Rule comprehension by study and task design*

		Rel. Freq (%)				Difference	
Item	Response	S1Ex	S2Re	S3Ex	S3Re	$\chi^2(df)$	$Pr(>\chi^2)$
Uniform	1,000	22.4	25.3	23.8	20.3	31.8(9)	< 0.001
	1,400	5.5	11.7	12.6	13.4		
	Equally likely*	61.6	52.6	54.0	55.9		
	Don't know	10.4	10.4	9.6	10.5		
Normal	*1,300**	66.0	55.2	64.3	59.9	29.1(9)	< 0.001
	1,500	3.2	8.2	7.0	4.9		
	Equally likely	28.9	32.9	26.2	31.6		
	Don't know	1.9	3.8	2.5	3.6		

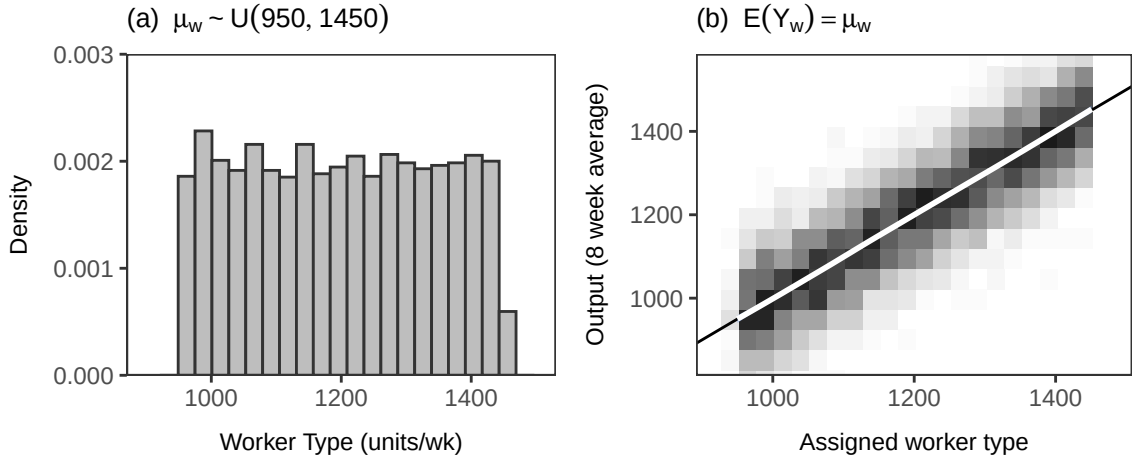
Note: Chi-squared tests of independence by item.
* Correct response.

3 Randomization checks

3.1 Assignment of worker type and weekly output

We randomly assign subjects *three* workers with a mean production, or type, μ_A, μ_B, μ_C,[1] drawn independently from a uniform distribution ranging from 950 to 1,450 units ($\mu_w \sim U(950, 1450)$). The weekly output for each worker, Y_w, follows a normal distribution with a mean equal to the assigned type and fixed standard deviation ($Y_w \sim N(\mu_w, \sigma)$). We used the JavaScript function Math.random() to draw worker types and the Box-Muller algorithm to generate weekly payouts. Figure 1 assesses the success of our programming in assigning worker types.

Figure 1: *Randomization check: worker types*



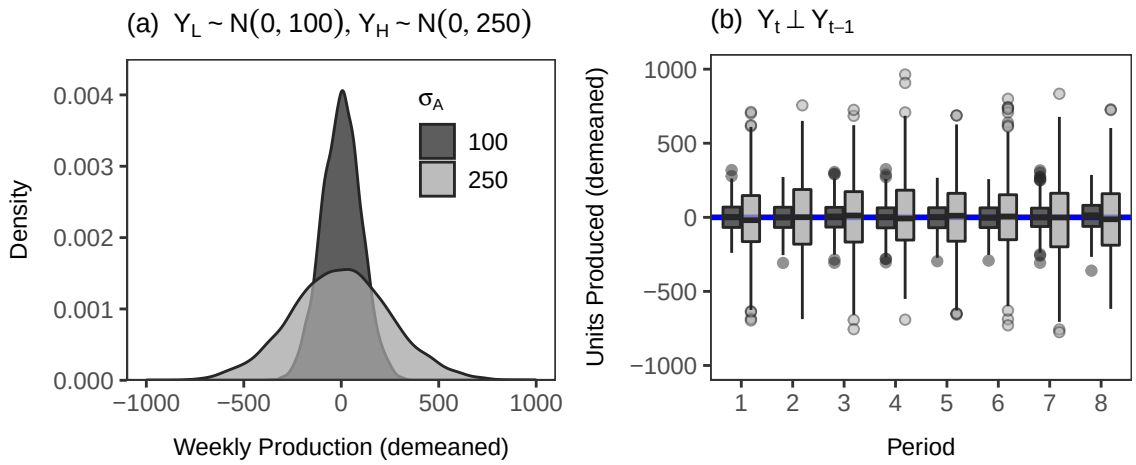
Note: Panel A is a histogram of the workers' assigned type. Panel B presents a "heatmap" of each worker's observed average production over assigned type. Darker areas indicate greater frequency. The diagonal line signifies equality between observed and assigned type. The white line is the lowess estimates with 95% confidence band.

The histogram in Panel A reveals the uniform distribution of worker type across all studies. As expected, the distribution ranges from 950 to 1,451 (rounding) with a mean 1,200 and median of 1,199. Panel B plots the workers' average observed output as a function of their assigned type. Darker colors in the heatmap indicate a higher frequency of participants. The diagonal line is a 45-degree line. The white loess estimate with a 95% confidence band reveals the fidelity between assigned and observed type. More specifically, the linear relationship is indistinguishable from the 45-degree line ($\bar{Y}_w = -5.7[9.9] + 1.00 * \mu_w[0.01]$, OLS standard errors in brackets).

Worker output in a given week is a random draw from a normal distribution – $Y_w \sim N(\mu_w, 250)$. The one exception is Worker A's production in the low noise condition of the Study 1 Extraction task: $Y_{A,low} \sim N(\mu_A, 100)$. Here we evaluate the fidelity of observed output to these parameters in the Study 1 Extraction task. Panel A of Figure 2 plots the density of A's weekly production (demeaned by A's type, μ_A) by experimental condition. The demeaned weekly yield among those in the low variance condition (gray) follows an approximately normal distribution with mean 1.61 and standard deviation of 99.8. The distribution of high variance yields has a mean equal to 1.1

[1] Worker C is the potential replacement for Worker B. C appears only if the respondent opts to replace Worker B with a anew worker

Figure 2: *Weekly output, Study 1 Extraction task*



Note: Panel A is a density plot of A's weekly production (demeaned). Dark and light gray coloring signify assignment to the low and high variance conditions respectively (data from the Study 1 extraction task only). Panel B is a box plot of Worker A's production (demeaned) by week. We limit the graph to weeks 1-8 for simplicity.

and standard deviation of 252. In all cases, these match the intended parameters.

Panel B of Figure 2 plots the demeaned yield by week for both experimental conditions. Indicative again of the success of the programming, the distributions are consistent in shape and statistical moments across weeks. There is no evidence of temporal dependence in worker yield.

3.2 Treatment balance

Table 3 evaluates demographic balance by assignment to experimental condition in the Extraction task (Study 1) and treatment arm in Study 3. One-way ANOVA estimates reveal a slight gender imbalance in Study 1, but otherwise no demographic imbalance in either study. The results also show no significant difference in rules comprehension by treatment group within these studies.

Table 3: *Demographic balance across treatments*

	Study 1 DV: Noise, σ_{Y_A}		Study 3 DV: Task Design	
Variable	$F(df)$	$Pr(>F)$	$F(df)$	$Pr(>F)$
Age	1.12(4)	0.35	0.70(4)	0.59
Education	0.99(6)	0.43	0.07(6)	0.99
Race	0.28(5)	0.93	0.55(5)	0.74
Female	12.38(1)	0.00	0.04(1)	0.85
Rules	0.73(2)	0.48	0.03(2)	0.97

Note: One-way ANOVA estimates by study.

4 Regression estimates and robustness checks

4.1 Estimates described in the main text

Tables 4 presents estimates of the performance voting response among subjects assigned to the Extraction and Recognition arms of Study 3 respectively. Specifically, they give OLS estimates of the choice to extend Worker B's contract as a function of worker types.

Table 4: *Performance voting, Study 3*

	DV: Extend the Target's contract	
	Extraction Arm	Recognition Arm
Target performance, β_T	0.106	0.137
	(0.013)	(0.018)
Comparator performance, β_C	-0.073	-0.080
	(0.014)	(0.018)
Intercept	0.307	-0.043
	(0.231)	(0.329))
Hypothesis tests		
No performance vote, $P(\beta_T \leq 0)$	0.000	0.000
No benchmark vote, $P(\beta_C \geq 0)$	0.000	0.000
Observations	473	247
R^2	0.173	0.279

Note: OLS estimates with standard errors in parentheses. We scale performance measures in 100s of units to facilitate interpretation.