Supplemental Materials
Quality Control (2023)

Austin Hart* J. Scott Matthews †

March 22, 2023

Citation for original work:

Hart, A. and Matthews, J.S. (2023). *Quality Control: Experiments on the Microfoundations of Retrospective Voting* (Elements in Experimental Political Science). Cambridge: Cambridge University Press.

Summary

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

A Recruitment and sample demographics

A.1 Human subjects and pre-registration

The studies reported in this Element were conducted in compliance with relevant laws in the United States and Canada and were approved by the Institutional Review Board of American University and the Interdisciplinary Committee on Ethics in Human Research at Memorial University. We also pre-registered the design, hypotheses, and analysis plans for our experiments (except Experiments 5, 6, and 8, which were the first studies completed chronologically). Links to pre-registration are presented in the main text.

A.2 Recruitment on MTurk

We recruited subjects for this study using Amazon.com's Mechanical Turk (AMT) and Turk Prime (now CloudResearch) marketplaces. We offered AMT workers over the age of 18 and operating with a U.S. IP address $0.40 (up from $0.30 prior to 2022) 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 mitigate bias in the type of AMT worker responding to the request, we automated calls for batches of 9 to 100 respondents (depending on the study) at fixed intervals. We used an IP filter in Qualtrics and an embedded Java screen in AMT to prevent repeated participation. As of 2021, we also used an external proxy and VPN detection API to identify IPs outside the US, flagged or suspicious IPs, and anyone using proxy or VPN services to mask their location.

A.3 Sample characteristics and attentiveness

From summer 2018 to spring 2022 9,157 individuals followed the link to an experimental task, hosted in Qualtrics. 99% consented to participate. We removed about 16% of respondents for failing a screener question designed to limit "click-through" behavior. In total, 7,309 respondents completed an experiment. The median time to completion was 5 minutes, 18 seconds. Participants were 44% female and 70% white with a median age 30-39.

A.4 Payment

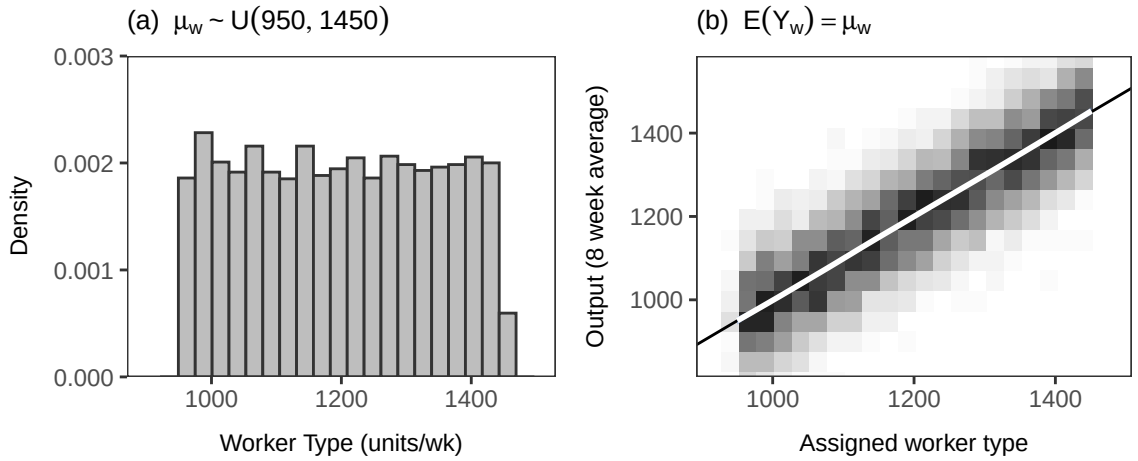
In addition to the base payment, participants earned bonus payments proportional to their workers' output over sixteen weeks. Given the uniform distribution of worker type and normal distribution of weekly output, the mathematical expectation (i.e. the average payout for participants evaluating the Incumbent worker at random) is \$0.67 in each task).

B Randomization checks

For most experiments, we randomly assign subjects to supervise workers with a mean production, or type, e.g., μ_w, 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 randomizing 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 higher frequency. 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 ($\bar{Y}_w = -4.7[6.7] + 1.00 * \mu_w[0.01]$, OLS standard errors in brackets).

C Game rules

We presented respondents with game 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, under 2% chose to review the rules. Here we present the text of the rules in each game design and special notes from individual experiments. Note that we name the workers Worker A and Worker B in the game. We refer to these workers as the Comparator and the Incumbent in the main text. For reference, Table 1 lists each experiment's purpose, design, and sample size.

Table 1: *Summary of experiments*

Exp	Performance Streams (n)	Underlying Design	Motivating questions
1	Single (248)	Baseline	Can voters integrate and appraise performance?
2	Single (403)	Baseline	Does performance variability inhibit performance voting?
3	Single (719)	Baseline	Is bad performance punished asymmetrically?
4	Single (764)	Baseline	Does timing (primacy/recency) matter?
5	Multiple (849)	Unmasking	Can voters unmask incumbent competence?
6	Multiple (368)	Recognition	Can voters recognize incumbent performance?
7	Multiple (550)	Unmasking	How do voters cope with an incongruent comparator?
8	Multiple (720)	Unm/Rec	Do voters want to benchmark?
9	Multiple (1,376)	Unm/Rec	Does information matter?
10	Single (394)	Baseline	Do voters choose quality benchmarks?
11	Multiple (918)	Hybrid	Can voters recognize (and adjust for) bad benchmarks?

C.1 Rules, baseline game

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.

C.2 Rules, unmasking game

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.

C.3 Rules, recognition game

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.

C.4 Rules amendment, Experiment 3)

Experiment 3 explores asymmetry in response to negative versus positive performance. We assign subjects to one of three fixed worker profiles: negative, neutral, or positive. These pro-

files were set in advance. To highlight their deviations from average/expected performance, we highlight the midpoint of worker types in the instructions. Specifically, we add to screen 1 of 2 of the baseline instructions the following: "By this standard, an above average worker typically produces more than 1,200 units per week while a below-average worker typically produces less than 1,200 units per week."

C.5 Rules amendment, Experiment 7)

Experiment 7 is multi-worker extension of the Unmasking design. Rather than a single worker, Worker A, operating as an exogenous distraction, Experiment 3 uses two workers, A and B, operating jointly as the disturbance. Respondents observe their joint production for the first eight weeks. Worker C, the target or Incumbent worker, arrives in week 9, when subjects then view the total factory output before voting to reappoint or replace the Incumbent. The rules, then, follow the Unmasking rules with the slight adjustment for three rather than two workers.

C.6 Offered performance cues, Experiment 8

Note that Experiment 8 assigns participants at random to complete either the Unmasking or Recognition task. As such, the rules in each condition are presented as above. The one difference 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

A respondent who selects "B's average production" and "Comparison of A and B's productivity," for instance, would then view the following:

Selected performance reports

REPORT: Worker B's average production

Worker B produced, on average, [$\bar{B}_{9:16}$] units.

REPORT: Comparison of A and B's productivity

On average, Worker B produced [$|\bar{B}_{9:16} - \bar{A}_{1:8}|$] units per week [*less/more*] than Worker A.

C.7 Assigned performance cues, Experiment 9

Note that Experiment 9 replicates Experiment 8 with the key exception that we assign subjects at random to view one of three performance reports: (i) Worker B's average production, (ii) the difference between B and A's average production, (iii) bonus earned to date. We present the assigned report as above, and use a five-second timer to encourage exposure to and engagement with the treatment report.

C.8 Rules amendment and cue, Experiment 10)

Experiment 10 builds on the baseline game design to explore the kinds of benchmark cues that individuals prefer. For this study, we remove explicit references to the parameters of the uniform distribution of worker types. With this in mind, we amend the "About the workers" instructions to the following:

About the Worker

You just transferred an employee, worker A, from another department. While it is not yet clear how A will perform, worker A is fully trained to work in any department.

Note that an employee's production varies each week for reasons beyond their control. Whether the worker is above or below average, weekly output tends to follow a normal (bell-shaped) distribution. This means that weekly output tends to balance around the worker's average, and exceptional output

above or below this average is uncommon.

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.

We then offer a choice among performance cues. The question, shown below, appears immediately after week 8 or week 16.

As you evaluate Worker A, you may wish to review factory performance records for reference.

You may select one (1) of the reports below, giving the average weekly output for different workers in your department in recent years.

- Lowest producing worker
- Worker in the bottom 25%
- Average worker in your department
- Worker in the top 25%
- Highest producing worker

C.9 Rules amendment, Experiment 11

Experiment 11 is a multi-stream game that combines elements from several prior designs. As with Experiment 10, we remove explicit references to the normal distribution of worker types. And as with the Unmasking design, we introduce a referent, or comparator worker. The "About the worker" screens, then, are given as follows:

Your Employees: Worker **A**

Worker **A** just transferred from another department. While it is not yet clear how **A** will perform, **A** is fully trained to work in any department.

Note that an employee's production varies each week for reasons beyond their control. Whether the worker is above or below average, weekly output follows a normal (bell-shaped) distribution. This means that weekly output tends to balance around the worker's average, and exceptional outputs above or below that average are uncommon.

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.

Your Employees: Worker **B**

As you evaluate worker **A**, you may wish to consider worker **B**'s performance for comparison.

Worker **B** is a longtime employee and, based on factory records, an average performer. About half of past employees were more capable than **B** and about half were less capable.

[comparable condition cue] **A** and **B** work independently, and they operate comparable machines.

[less efficient condition cue] **A** and **B** work independently. However, with the new hire, *worker **B** will move to a machine that is considerably slower and less efficient than **A**'s.*

[more efficient condition cue] **A** and **B** work independently. However, with the new hire, *worker **B** will move to a machine that is considerably faster and more efficient than **A**'s.*

Note that Experiment 11 participants are also assigned at random to one of three cues about Worker B's machine. See the box above for full text.

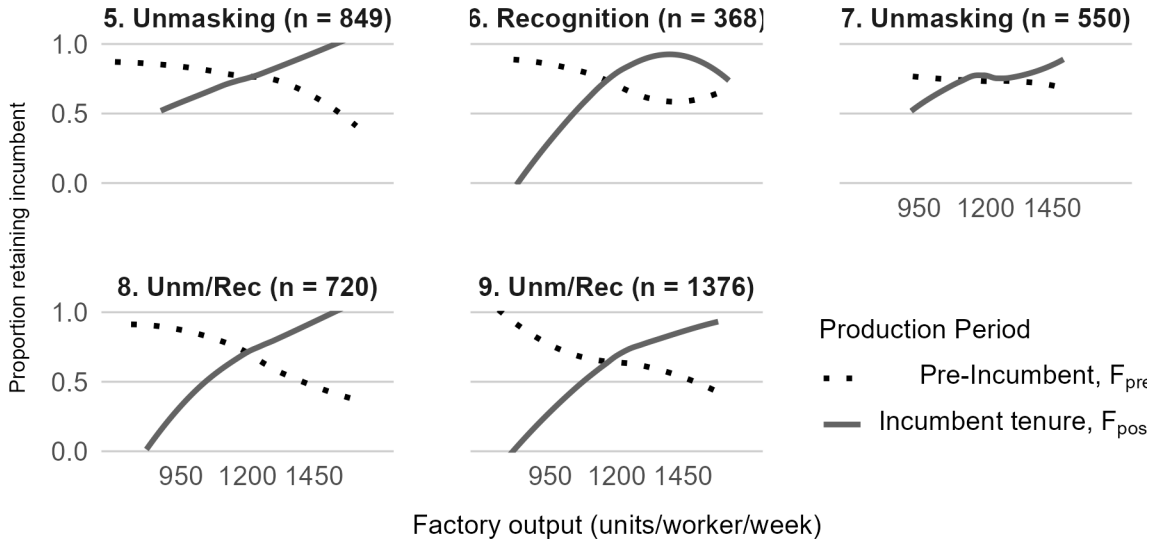
D Model estimates and alternate specification, Eq. 1

Figure 2 plots loess estimates of the Incumbent's reappointment rate as a function of factory performance before and during the Incumbent's tenure ($F_{pre,i}$ and $F_{post,i}$, respectively). Note that this is only possible for Experiments 5-9, wherein we randomize the type, or underlying competence, for both Incumbent and Comparator workers.

Table 2 presents OLS estimates of Equation 1 from the main text regressing the decision to reappoint the Incumbent on the effect of factory performance. The results show that the probability of reappointing the Incumbent rises with performance during the Incumbent's tenure (i.e. $\hat{\beta} > 0$ in all studies). However, net of the factory output during the Incumbent's tenure, the probability of reappointment declines with factory performance prior to the Incumbent's arrival (i.e. $\hat{\alpha} < 0$ in all studies). All estimates are statistically significant.

Given our study design, and in light of Arel-Bundock et al.'s (2021) critique of prior benchmarking research, we argue that these are the most appropriate models for distin-

Figure 2: *Baseline response to production, multi-stream games*



Note: Loess estimates of the proportion of respondents opting to reappoint the Incumbent by average factory output before and during the Incumbent's tenure. We scale output by number of workers for comparability.

guishing blind retrospection, rational discounting, and benchmark processing. Critically, they mimic the performance voting dilemma that voters face in the real world. However, some readers may prefer to model Incumbent support as a function of the workers' respective types rather than observed performance. Note also that our analysis plan for Experiment 7 referred to this approach. Table 3 presents these results. Overall, the results are robust to this alternative specification. The one difference is the response to the Comparator's type in Experiment 7 (the multi-worker disturbance game). Though in the same direction as the estimated response to observed factory production in the main specification, the effect is insignificant and we cannot reject a null hypothesis of no benchmarking. As we posit in the main text, participants turning away from an incongruous point of comparison is not inconsistent with our main and persistent finding of benchmark processing. In the presence of a congruous—though irrelevant—Comparator, participants in four experiments exhibit behavior consistent with this alternative approach to modeling the benchmark response.

Table 2: *Response to factory performance, multi-stream games*

	Factory Performance					
	Pre-Incumbent ($F_{pre,i}$)		With Incumbent ($F_{post,i}$)		Intercept	
Experiment (n)	$\hat{\alpha}$	$Pr(\alpha \leq 0)$	$\hat{\beta}$	$Pr(\beta \geq 0)$	$\hat{\theta}$	se
5. Unmasking (849)	-0.093	0.000	0.135	0.000	0.251	0.145
6. Recognition (368)	-0.055	0.000	0.123	0.000	-0.105	0.214
7. Unmasking (550)	-0.114	0.000	0.158	0.000	0.211	0.191
8. Unmasking (473)	-0.129	0.000	0.151	0.000	0.437	0.196
8. Recognition (247)	-0.065	0.000	0.153	0.000	-0.410	0.268
9. Unmasking (894)	-0.138	0.000	0.192	0.000	0.002	0.153
9. Recognition (482)	-0.047	0.000	0.156	0.000	-0.695	0.190

Note: OLS estimates of seven models regressing the vote to reappoint the Incumbent on observed factory performance. We scale types in 100s of units to facilitate interpretation.

Table 3: *Response to assigned worker types by study*

	Incumbent (μ_I)		Comparator (μ_C)		Intercept	
Experiment (n)	$\hat{\beta}_I$	$Pr(\beta_I \leq 0)$	$\hat{\beta}_C$	$Pr(\beta_C \geq 0)$	$\hat{\theta}$	se
5. Unmasking (849)	0.076	0.000	-0.031	0.001	0.225	0.165
6. Recognition (368)	0.123	0.000	-0.058	0.000	0.118	0.208
7. Unmasking (550)	0.064	0.000	-0.011	0.187	0.118	0.208
8. Unmasking (473)	0.106	0.000	-0.073	0.000	0.307	0.231
8. Recognition (247)	0.137	0.000	-0.080	0.000	-0.043	0.329
9. Unmasking (894)	0.107	0.000	-0.045	0.000	-0.084	0.170
9. Recognition (482)	0.156	0.000	-0.047	0.000	-0.702	0.236

Note: OLS estimates of seven models regressing the vote to reappoint the Incumbent on workers' underlying competence, or types. We scale types in 100s of units to facilitate interpretation.