Sequential Sampling by Individuals and Groups: An Experimental Study

By

Pellumb Reshidi, Duke University
Alessandro Lizzeri, Princeton University
Leeat Yariv, Princeton University
Jimmy Chan, National Taiwan University
Wing Suen, University of Hong Kong

Griswold Center for Economic Policy Studies
Working Paper No. 312, October 2022
Revised version, May 2024

Sequential Sampling by Individuals and Groups: An Experimental Study*

Pëllumb Reshidi† Alessandro Lizzeri‡ Leeat Yariv§ Jimmy Chan¶ Wing Suen‖

May 17, 2024

Abstract

Many committees—juries, political task forces, etc.—spend time gathering costly information before reaching a decision. We report results from lab experiments focused on such dynamic information-collection processes, as in sequential hypothesis testing. We consider decisions governed by individuals and groups and compare how voting rules affect outcomes. Several insights emerge. First, average decision accuracies approximate those predicted theoretically, but these accuracies decline over time: participants display non-stationary behavior. Second, groups generate markedly different outcomes than individuals, with majority rule yielding faster and less accurate decisions. In particular, welfare is higher when sequential information is collected in groups using unanimity.

*A previous version of the paper was circulated with the title "Individual and Collective Information Acquisition: An Experimental Study." We thank Marina Agranov, Roland Benabou, Daniel Benjamin, Tim Cason, Stephen Morris, Salvatore Nunnari, Wolfgang Pesendorfer, the Editor Dirk Bergemann, and six anonymous reviewers for very helpful discussions and feedback. We gratefully acknowledge the support of NSF grants SES-1629613 and SES-1949381, as well as award NTU-112V1027-2 from Taiwan's Ministry of Education's Yushan Fellow Program.

†Department of Economics, Duke University pellumb.reshidi@duke.edu

‡Department of Economics, Princeton University lizzeri@princeton.edu

§Department of Economics, Princeton University lyariv@princeton.edu

¶Department of Economics, National Taiwan University jimmyhingchan@ntu.edu.tw

‖Faculty of Business and Economics, University of Hong Kong wsuen@econ.hku.hk

1 Introduction

1.1 Overview

Information acquisition precedes a variety of important decisions. The pioneering work of Wald (1947) illustrated that collecting costly information dynamically, using *sequential hypothesis testing*, is more efficient than deciding statically on the sample size, using *classical hypothesis testing*.[1] In many settings, information is indeed collected sequentially. Agencies like the FDA and the EPA follow sequential evaluations, with each stage contingent on the one preceding it; juries recall testimonies through dynamic deliberations; hiring committees often proceed via several interview stages. In addition, many important information-collection decisions are made by groups: expert committees, juries, or department faculty. The theoretically optimal dynamic policy requires monitoring of information and choices at every point in time. In real world settings, it may be too complex to deliver its theoretical advantages. In practice, does sequential sampling afford the efficiency benefits predicted theoretically? Do groups behave differently from individuals? Does the aggregation procedure within groups matter? There is a dearth of data addressing these questions. Our aim is to fill this gap.

This paper reports results from lab experiments inspecting sequential information collection by individuals and groups, where we consider groups operating under different aggregation procedures. In our design, participants trade off information accuracy and costs dynamically. In applications, groups may exhibit a multitude of context-specific features: preference heterogeneity, available communication channels, etc. Our goal is to take an initial step by examining group interactions without considering motives related to preference aggregation or strategic information sharing. In our design, participants have identical incentives and information across all treatments. Theoretically, optimal choices are independent of the aggregation rule. In particular, individuals and groups should use a stationary cutoff posterior, as in Wald (1947), to determine when information collection comes to a halt.

Our data reveal the following patterns. First, on average, decision accuracies come close to those predicted by the theory, but these accuracies decline over deliberation time. That is, partici-

[1] Wald (1947)'s insight was perceived so profound that, in the introduction to his book, it is noted that "Because of the usefulness of the sequential probability ratio test in development work on military and naval equipment, it was classified Restricted within the meaning of the Espionage Act."

pants display time-declining thresholds instead of the theoretically predicted stationary thresholds. Second, groups induce markedly different outcomes than individuals, with majority rule leading to faster and less accurate decisions. In fact, welfare is highest when information is collected in groups using unanimity.

Our results have implications for institutional design when information collection is an important component of decision making. Using committees, rather than individuals, can be beneficial even if increasing the size of the decision body does not affect the overall information available. However, this benefit is only attained when more stringent rules govern deliberation.

Our design implements a sequential sampling setup.[2] There are equally likely states, A or B. Ultimately, each participant must guess the state of the world and gets rewarded when correct. Each state is associated with a Brownian motion. The drift is μ when the state is A and $-\mu$ when the state is B. The Brownian motion's variance is state-independent. As time goes by, the realized sample path of the Brownian motion becomes increasingly informative about the underlying state. To avoid confounding our results with updating biases, at every instant, participants observe the posterior probability of each state, rather than the underlying Brownian path. There is a flow cost of information collection. Whenever information collection terminates, participants submit their guesses. Our focus is on the trade-off pertaining to information collection: waiting longer before making a decision increases accuracy, but comes at a cost. In our benchmark treatments, decisions are made by individuals, as in the canonical paradigm. In additional treatments, decisions are made in groups. When in a group, we consider two commonly used institutions: majority and unanimity. Group members decide whether to stop or continue information collection at each point in time. Under majority, whenever two members agree on a guess, information collection terminates for the group, and the majority guess is submitted. Analogously, under unanimity, whenever all members agree on the guess, information collection terminates, and that guess is implemented. In our group treatments, information is public: group members are privy to the same information. Furthermore, group members receive the same payoff, derived from the common costs accrued during the group's information-collection period and the group's guess accuracy.

Theoretically, it is optimal to stop once there is sufficient confidence in guessing the state, namely when the posterior belief exceeds a time-independent threshold: at that point, the costs of

[2] Specifically, our setup mimics that of Dvoretzky et al. (1953).

additional information surpass its benefits. With our parameters, the optimal threshold posterior is 0.81. In our experimental treatments, individuals' mean posterior at decision time is relatively close to that predicted by theory, standing at 0.77. However, participants use time-dependent thresholds that decrease over time.[3] In particular, contrary to theoretical predictions, quicker decisions tend to be more accurate. As we discuss in our literature review, this observation is consistent with a wide neuroscience literature documenting a similar pattern using perception tasks. Our data are unique since, by design, we directly observe the posterior probabilities participants see over time.

Theoretically, in our treatments, efficient group outcomes should coincide with individual outcomes. Nonetheless, in our data, group outcomes significantly differ from individual outcomes, and the difference depends on the voting rule in place.

In groups using unanimity, members behave as in the individual treatment. Heterogeneity in individual behavior aggregates into more accurate decisions by such groups: the most demanding agent in a group governs decisions. In contrast, under majority, participants make significantly quicker decisions relative to our individual treatment, and group choices are significantly less accurate. Welfare is then maximal when groups operate under unanimity. Furthermore, welfare is higher in all of our treatments than the welfare were agents choosing the optimal duration of information collection at the outset, as in classical hypothesis testing.

It is challenging to find a force rooted in group dynamics that explains hastiness under majority but has no effect under unanimity. We suggest individuals may display a *demand for agency*: a desire to have a say in the vote. Under unanimity, all group members express their preferences before a decision is made and a demand for agency is moot. Under majority, a decision can be made before all members have cast their vote, and a demand for agency introduces a potential race. Another possible explanation relates to our finding of declining decision thresholds. Under majority, best responses to other group members' decreasing thresholds involve faster decisions. Under unanimity, best responses remain the individually optimal thresholds.

[3] This pattern does not extend across rounds: participants' mean stopping posteriors are higher in the second half of our sessions.

1.2 Related Literature

Sequential sampling, proposed by Wald (1945, 1947), introduced the idea of collecting data sequentially to improve the efficiency delivered in classical, static, hypothesis testing. With each piece of data, a likelihood ratio test is performed to determine whether more observations are needed to accomplish a desired level of statistical confidence. Sequential sampling has been used widely to describe how individuals collect information and to guide researchers in data collection; see Dominitz and Manski (2017) and references therein.

Strulovici (2010), Chan et al. (2018), and Henry and Ottaviani (2019) consider environments in which information collection is sequential: the committee decides at each date whether to continue acquiring costly information, or stop and choose an alternative. In particular, Chan et al. (2018), which our group treatments mimic, as well as Henry and Ottaviani (2019) and McClellan (2021), build on the literature on sequential hypothesis testing that started with Wald (1947).

A large experimental literature studies how individuals collect and process information statically (Benjamin, 2019). There is also recent work that studies static information collection in groups; see Grosser and Seebauer (2016), Bhattacharya et al. (2017), Ginzburg and Guerra (2019), and Elbittar et al. (2020).

An old experimental literature in psychology considers dynamic settings similar to ours, looking at individual choices; see Becker (1958), Edwards and Slovic (1965), Green et al. (1964), and Fried and Peterson (1969). We contribute in several ways. First, we investigate how *groups* using different rules collect information. Methodologically, our design provides participants with evolving posteriors and is not confounded with updating heuristics. Last, our participants can collect an effectively unbounded—rather than finite—volume of information, in line with the sequential sampling model.

Several papers inspect individual dynamic search behavior experimentally, see Gabaix et al. (2006), Brown et al. (2011), Caplin et al. (2011), and references therein. In these experiments, participants spend resources dynamically in order to identify a good alternative, but the underlying optimization problem differs from ours. Chen and Heese (2021)'s experiment resembles our individual treatment, but focuses on alternatives' ethical valence.[4]

The neuroscience literature has produced a rich body of work that inspects binary perceptual

[4]Relatedly, Freer et al. (2020) test a 3-period version of collective experimentation via groups.

tasks. Response times are often interpreted as costly, turning the problem into a sequential sampling one, often termed the drift-diffusion model. See, for instance, Swensson (1972), Luce et al. (1986), Ratcliff and Smith (2004), and Ratcliff and McKoon (2008). The main finding emerging from this literature is that quick decisions tend to be more accurate. This insight aligns with our observation of declining thresholds: as time passes, our participants stop information collection with less certainty on the correct choice. An important contrast with these studies is that we provide the posterior probability that any choice is correct over time. This allows us to speak directly to new theories of dynamic choice that have appeared recently, see Baldassi et al. (2020) and Fudenberg et al. (2018).

2 Experimental Design

A description of the interface and sample instructions are available in the Online Appendix. At the core of our experimental design is the choice of the amount of information to acquire prior to making a binary decision. There are two possible states: A and B. Although labeled neutrally in the lab, these can stand for a guilty or innocent defendant in the jury context, a good or bad policy in the political context, a profitable or unprofitable investment in a finance context, etc. At the start of each period, one state is chosen randomly with probability $1/2$. Participants ultimately need to guess the state and are paid based on the correctness of their guess. In the lab, participants receive \$2 for a correct guess and nothing otherwise.

Before making their guess, participants have access to information that evolves according to a continuous-time Wiener process. The process has state-independent instantaneous variance σ^2, but state-dependent drift. When the state is A, the drift is μ; when the state is B, the drift is $-\mu$. To produce reasonable expected round durations throughout our treatments, we set $\mu = 0.84$ and $\sigma^2 = 1$. Naturally, our experimental software provides an approximation of the continuous setup: the interface is updated five times a second.[5]

The evolution of the realized Wiener process provides continuous information on the likelihood of either state. Nonetheless, the Bayesian calculus necessary to deduce this likelihood is non-trivial.

[5] We use a single parameterization, which allows for a rich study of information collection in a variety of settings with a manageable number of treatments. Our design is amenable to different parameter choices, offering a natural future step.

The difficulty this calculus introduces is orthogonal to our investigation.[6] To mitigate the impacts of participants' limitations in statistical analysis, our design directly displays the evolution of the *probability* that the state is A (or B), using a uni-dimensional scale depicting the moving posteriors. Furthermore, this design element echoes features of a variety of applications, where decision makers receive digested information that they base their choices on: the FDA, politicians, boards of trusties, often receive summary results of various studies, rather than analyze the underlying data themselves.

Our treatments vary whether choices are made by individuals, groups using majority rule, or groups using unanimity rule. Each participant took part in only one of the three treatments. We now describe their details.

Our treatments mimic the sequential-sampling environment of Dvoretzky et al. (1953). Participants observe information evolve over time and, at each instant, can guess A, B, or wait for further information by choosing W. Information comes at a flow cost of 40 cents a minute.

In the treatment in which individuals make decisions on their own—the *individual* treatment—a round ends as soon as a participant makes an A or B guess.

In our group treatments, participants are randomly matched to form groups of three in each round. Information is public: all individuals in the group observe the same information. A round ends as soon as a quorum of q individuals agrees on an A or B guess. In the *majority* treatment, $q = 2$. In the *unanimity* treatment, $q = 3$. As long as a quorum has not been reached, participants can change their decisions between A, B, and W at any time. Throughout, participants can see the choices of other group members.

Our parameters are chosen so that the optimal stopping threshold is not too extreme and errors are unlikely to be one-sided. Specifically, in our individual treatment, a risk neutral agent should wait until the probability of the most likely state is 0.81. The utilitarian efficient equilibrium for groups, under both majority and unanimity, corresponds to the same policy, with each group member utilizing a posterior threshold of 0.81. See the Online Appendix for details.

Feedback and Payments In all treatments, the feedback at the end of each round contains participants' payoffs. In groups, all members are paid the same amount, incorporating the correctness

[6] It is well known that lab participants are frequently challenged by statistical updating; see, for instance, the survey of Benjamin (2019).

of the group's guess and information costs.

Each treatment was preceded by two practice rounds, followed by 30 payoff-relevant rounds. Participants were ultimately paid for 20 randomly selected rounds out of these 30.

Information Processes The 30 information processes participants experienced in the experimental rounds were identical across treatments. To select these processes, we randomly generated 15 sample paths with the parameters specified above. These processes are "representative" in that the mean, median, and five quintiles of the theoretically optimal sequential stopping times match those of the underlying distribution. These processes correspond to the first 15 real rounds in each treatment. The last 15 processes in each treatment were derived by generating the reflected "mirror images" of the first 15 processes. Namely, whenever the realized state in the original process is A (or B), it is B (or A) in the reflected process. Furthermore, at any instant, if the original process indicates a probability p that the state is A, the reflected process indicates a probability $1-p$ that the state is A. The reflected processes were used in the same order as the original processes. In that way, participants effectively faced the same 15 decision problems twice during a session. This design element allows us to evaluate learning in a highly controlled fashion.[7]

Auxiliary Elicitations At the end of each session, participants completed two risk-elicitation tasks as in Gneezy and Potters (1997) and two dictator games, with various different parameters. Participants were paid for one randomly-chosen risk-elicitation task and one randomly-chosen dictator game.[8]

Summary Experiments were run at the Princeton Experimental Laboratory for the Social Sciences (PExL) with 130 Princeton undergraduate students as participants. We conducted four sessions for each group treatment, with 12 participants in each. We had 34 participants in our individual treatments. The experimental software was programmed using oTree (Chen et al., 2016).

Approach to Data Analysis Participants' behavior changes somewhat during early experimental rounds as they learn about the problem. We see no evidence for substantial learning in

[7]Identifying repetitions is extremely unlikely: it would require memorizing many ordered values and realizing they are mirrored.

[8]We elicited duplicate responses to allow for measurement-error correction as suggested in Gillen et al. (2019).

Table 1: Aggregate Behavior

	Mean Posterior		Mean Time Waited	
	All Rounds	Last 15	All Rounds	Last 15
Individual	0.77	0.78	33.56	37.55
	(0.003)	(0.005)	(0.687)	(1.12)
Majority	0.73	0.73	23.07	24.38
	(0.002)	(0.003)	(0.335)	(0.51)
Unanimity	0.82	0.84	46.71	53.68
	(0.002)	(0.003)	(0.724)	(1.11)
Theory	0.81		39.03	

Standard errors in parentheses

Notes: We report the mean posteriors and the mean time waited to reach these posteriors for the pivotal votes across treatments. Additionally, we provide the theoretically optimal posterior (0.81) and the expected time to achieve this posterior (39.03).

later rounds.[9] Throughout the paper, we present figures aggregated across all experimental rounds, as those displayed appear virtually identical when we use either the full data or the last half of our sessions. Regression results using the last 15 rounds appear in the Online Appendix. Our qualitative results remain the same. We see little variability across sessions and limited effects of group partners in early rounds on behavior in later rounds, which is why we use individual-level clustering. Results remain qualitatively similar with session-level clustering. Risk attitudes and altruism proclivities do not appear to play an important role in organizing our data, even after measurement-error correction. We, therefore, do not include data from these elicitations in our main specifications. See the Online Appendix for analyses documenting these observations.

3 Broad Patterns of Behavior

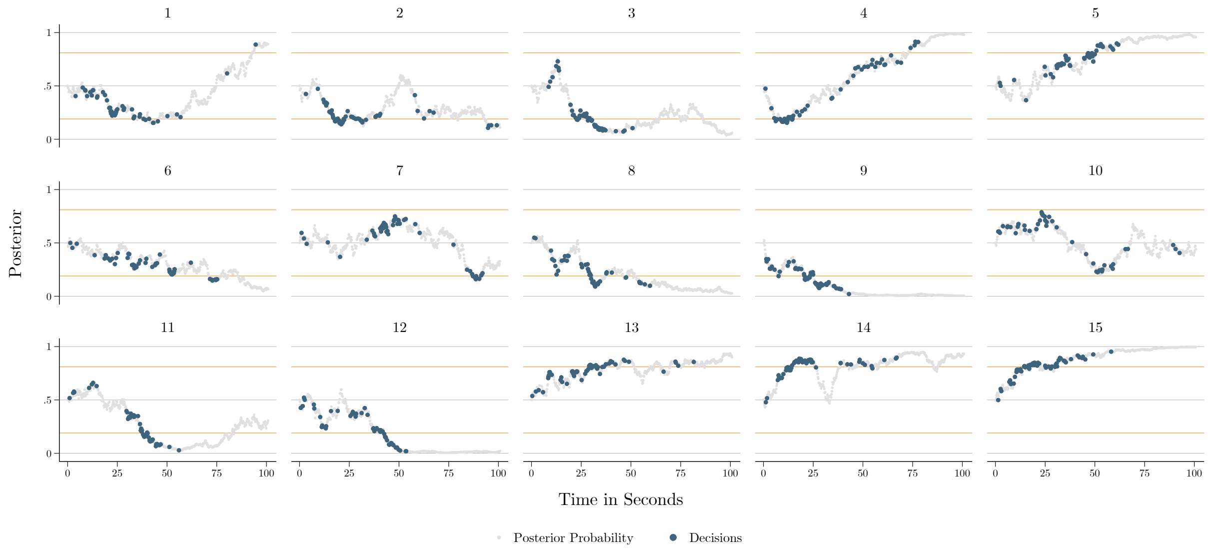
Our individual and majority treatments lead to less accurate decisions than theoretically predicted, whereas the unanimity treatment yields outcomes that are extremely close to those theory predicts, as seen in Table 1.

Differences between empirical threshold posteriors and those predicted by theory may, at first blush, appear small. Nonetheless, these differences translate to substantial differences in wait times. For instance, the unanimity treatment leads to double the wait time relative to our majority treatment. This is a direct consequence of the convexity of precision costs: the marginal time required to attain a given precision increases as precision rises.

Figure 1 depicts the evolution of posteriors and the choices made in each of our 15 processes in

[9] Recall that the processes utilized in the first and second half of each session are identical. In the Online Appendix, we compare equivalent round pairs to assess experience effects.

Figure 1: Pulling the Trigger: Individual Treatment



Notes: We report the evolution of posteriors (gray dots) across the 15 rounds we utilize. Blue dots represent the posteriors participants saw at the moment of casting their vote. Orange lines indicate the optimal thresholds (0.19 and 0.81). As processes are repeated in the first and last 15 rounds, each panel displays pooled data from the two rounds in which the corresponding process was used.

the individual treatment. In order to simplify the presentation, each panel aggregates observations from two reflected processes (for example, panel 1 corresponds to the first and sixteenth process, panel 2 to the second and seventeenth process, etc.). The Figure illustrates the point at which individuals "pulled the trigger" and selected an alternative.

The Figure suggests themes appearing in our more detailed analysis below. First, decisions are heterogeneous. Second, many participants stop information collection at the theoretically predicted posterior accuracy (corresponding to the horizontal orange lines within each panel). Furthermore, participants respond to information in that decisions are clustered around higher posteriors. Last, individuals become more lenient, requiring less accuracy to stop the longer they wait: they display decreasing thresholds. Consider, for example, process 10. Several individuals decide late in the process, when posteriors are close to 50%, despite choosing not to stop at earlier points when posteriors were close to 80%. The Online Appendix shows the analogous figure for our majority and unanimity treatments. Results are similar: we see more leniency over time.

Table 2: Probability of Voting - Probit Regression

	$P(Vote)$					
	Individual	Majority	Unanimity	Individual	Majority	Unanimity
Posterior	5.357***	5.149***	5.690***	5.071***	3.787***	5.463***
	(0.400)	(0.426)	(0.406)	(0.402)	(0.478)	(0.427)
Time	0.242**	0.798***	0.333***	0.313***	0.673***	0.328***
	(0.120)	(0.179)	(0.111)	(0.109)	(0.189)	(0.110)
Slope				0.137***	0.132***	0.0475*
				(0.0360)	(0.0338)	(0.0253)
Standard Dev				-0.142	0.626**	0.350*
				(0.212)	(0.275)	(0.203)
Constant	-4.980***	-4.626***	-5.263***	-4.891***	-3.880***	-5.192***
	(0.344)	(0.291)	(0.312)	(0.332)	(0.367)	(0.335)
N	7865	6772	11113	6824	5301	9660

Standard errors in parentheses
Individual-level clustering
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: For each treatment, we run a probit regression of the probability of casting a vote on the absolute difference between the posterior and 0.5 (denoted Posterior), the time in minutes within a round, the slope and standard deviation within 5-second non-overlapping windows, as well as a constant.

4 Behavior across Treatments

4.1 Individual Behavior and Non-stationarity

Theoretically, the probability of voting should be 1 when the posterior reaches its theoretical threshold value of 0.81, and 0 for any lower posterior. In particular, the probability of voting should respond only to observed posterior probabilities, not to the time that has passed; to features of the sample path; or to choices of other group members, if in a group.

We now assess the determinants of stopping behavior by using a probit regression whose coefficient estimates appear in Table 2. The left-hand side variable captures whether a participant has voted and the main explanatory variable is the posterior for the favored alternative. In the left panel of Table 2, we also include the time (in minutes) to allow for time dependence. To allow for the possibility of path dependence, in the right panel of Table 2, we include features of the sample paths. Specifically, we divide each round into (non-overlapping) 5-second time intervals. An observation corresponds to each such window and consists of whether a vote was cast, the posterior and time at the end of the window, as well as the slope and standard deviation of the sample path within the window.[10] We utilize data up until participants cast their individual votes.

The left panel of Table 2 indicates that decisions to cast a vote are responsive to posteriors, with higher posteriors naturally leading to increased voting probabilities. In addition, *stopping*

[10]The slope corresponds to the average posterior gain per minute calculated using the 5-second window. To make the posteriors and standard deviations ranges comparable across 5-second windows, we multiply the standard deviation by five. Different time windows, of 3,...,7 seconds, yield similar results. See the Online Appendix for details.

decisions are not stationary. Controlling for posteriors, the more time passes, the more likely agents are to make a decision. For example, in the individual treatment, waiting for one additional minute is equivalent to an approximate increase of 4.5 percentage points in the observed posteriors ($0.045 \times 5.357 \approx 0.242$). Furthermore, time appears to have a stronger impact on the likelihood of making a decision when groups use majority rule.

When including features of the sample paths, coefficients corresponding to posteriors and time passed change only slightly. However, the coefficient corresponding to *Slope* is positive and significant. That is, after a brief period in which posteriors increase rapidly, a decision is more likely, particularly for individuals and groups using majority rule. In contrast, the coefficient corresponding to *Standard Dev* is barely significant: recent variation in posteriors has a limited effect.

In the Online Appendix, we show that results remain qualitatively similar when clustering errors more conservatively at the session level. Similar conclusions emerge when we consider an alternative approach in which we analyze observed stopping posteriors directly.

Our non-stationarity results are difficult to reconcile with existing theories. For example, the literature relating to drift-diffusion models (DDM) also finds that quick decisions tend to be more accurate; see, e.g., Swensson (1972), Luce et al. (1986), Ratcliff and Smith (2004), and Ratcliff and McKoon (2008). The recent explanation provided by Fudenberg et al. (2018) for the relationship between speed and accuracy relies on decision-makers' uncertainty about payoffs, which translates into uncertainty about the process and leads to optimal non-stationary behavior. This explanation is challenging to apply to our data since participants have common knowledge of the payoff of each state. In our data, experience does not significantly reduce the degree to which thresholds are decreasing, suggesting that it is unlikely that "subjective uncertainty" about the process is what drives behavior. Furthermore, as shown, stopping behavior responds to the sample path itself, behavior that cannot be explained with a pre-determined (potentially time-variant) threshold, as in Fudenberg et al. (2018).[11]

Brown et al. (2011) experimentally study a stationary job-search problem with a known distribution of wage offers. They find that reported reservation wages decrease over time. They consider two potential explanations: non-stationary time discounting and a "duration discourage-

[11] Baldassi et al. (2020) offer an axiomatic approach to the drift-diffusion model. McClellan (2021) derives non-stationary threshold posteriors as the consequence of agency frictions. Strack and Viefers (2021) report on path dependence in a related search setting. See also references therein.

ment effect," whereby agents set reservation wages in response to cumulative costs. Both effects are present in their treatments, although the first is more pronounced. Certainly, convex costs of time, if calculated round by round, could yield declining thresholds. Interestingly, the average stopping durations increase *between* rounds, which is difficult to reconcile with participants having convex costs of time within a session.[12] Furthermore, as mentioned above, our participants also react to features of the process itself, a phenomenon that cannot be explained by the mechanisms suggested by Brown et al. (2011).

As mentioned above, our participants also react to features of the process itself, a phenomenon that cannot be explained by the mechanisms suggested by Brown et al. (2011). Furthermore, thresholds decline within a round but not within a session. In fact, the average stopping posteriors and durations are higher in the second half of the sessions.[13] Thus, convex costs would need to be narrowly framed for them to explain our data.

4.2 The Impacts of Decision Procedures

For each of our treatments, the left panel of Figure 2 displays the cumulative distribution functions (CDFs) of the threshold posteriors.[14] We see a substantial impact of the governing decision rule. Distributions can be ordered via first order stochastic dominance, with the unanimity treatment yielding the highest accuracy decisions and the majority treatment yielding the least accurate decisions. In particular, the averages presented in Table 1 are not principally driven by outliers.

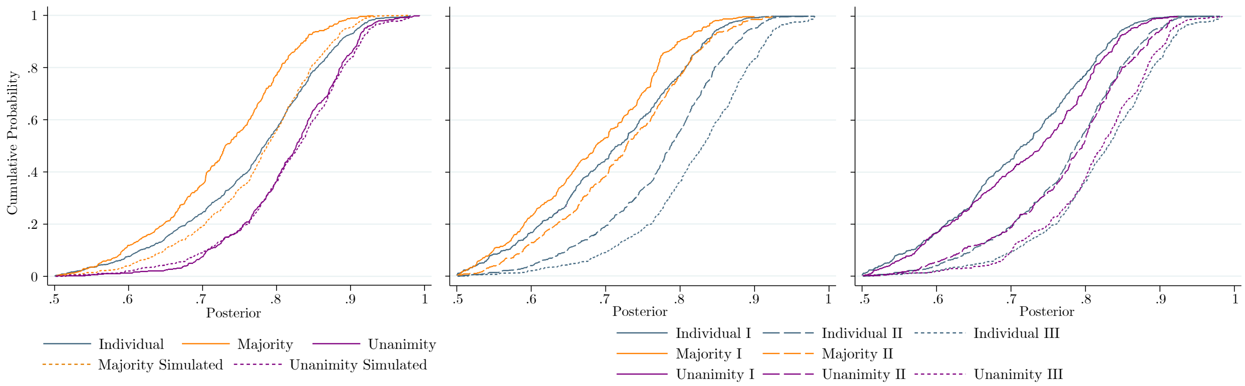
Theory suggests we should not observe any differences in outcomes among voting rules. However, theory also predicts a single threshold posterior, whereas we observe substantial heterogeneity in behavior in the individual treatment. We need to distinguish between behavior differences across treatments and outcome differences that are the consequences of aggregating three random and heterogeneous individuals. Specifically, groups governed by majority rule decide according to the second-order statistic, whereas groups governed by unanimity rule decide according to the third-order statistic. To assess whether treatment differences are purely due to aggregation, we simulate

[12]Since the second 15 rounds utilize the same processes as the first 15, the increase in the observed stopping posteriors cannot be an artifact of the processes' features.

[13]Since the second 15 rounds utilize the same processes as the first 15, the increase in the observed stopping posteriors cannot be an artifact of the processes' features.

[14]Participants vote for the alternative favored by the stopping posterior 96% of the time overall, and 98% of the time when pivotal.

Figure 2: Posterior CDFs and Vote Order



Notes: Left panel: we present the CDFs for each treatment alongside simulated CDFs for majority and unanimity based on the individual treatment. Middle and right panels: we compare the order statistics of simulated groups from the individual treatment with the observed order statistics from the majority and unanimity treatments, respectively.

hypothetical groups of three participants by drawing data from our individual treatment.[15] The left panel of Figure 2 presents the resulting CDFs from these simulated groups alongside the experimental distributions. The increased accuracy observed in groups using unanimity appears to be entirely due to aggregation. In contrast, groups using majority rule yield substantially less accurate and hastier decisions than those of simulated groups using majority, suggesting that hasty majority choices are not solely a consequence of an aggregation effect.[16] We discuss this phenomenon in more detail in Section 6.

The middle and right panel of Figure 2 presents the distributions of posteriors corresponding to the first, second, and third votes across our treatments. Alongside these distributions, we present analogous distributions for simulated groups of three generated from the individual treatment via the procedure described above.[17]

Individual voter behavior under unanimity is very similar to behavior of individuals deciding in isolation. This is true for the first, second, and third group members casting their votes. This observation reinforces the idea that outcome differences between the individual and unanimity treatments are exclusively due to the aggregation rule acting on heterogeneous individuals.

For our majority treatment, however, hasty behavior is not only a characteristic of the second

[15] For each round, we randomly group the 34 participants in our individual treatment into 11 groups of 3 participants, randomly discarding one. We do so $1,000$ times. Across all 30 rounds, $330,000$ groups are therefore simulated.

[16] A two-sided Kolmogorov-Smirnov test fails to reject the hypothesis that simulated and observed unanimity group decisions are identical. The same test rejects the hypothesis that simulated and observed majority group decisions are identical. Additional analysis in the Online Appendix illustrates that our results are robust to correlations across observations.

[17] Since a third vote is available in fewer than 10% of groups operating under majority, we do not present the third vote CDF here.

(and pivotal) voter; the first voter appears to be hasty as well. Both the first- and second-order statistics from the simulated individual treatment stochastically dominate the observed distributions corresponding to the first and second voters from the majority treatment. Interestingly, the distribution of second voters under majority is very similar to the distribution of first voters in the individual simulated treatment, a point we soon return to.

5 Performance

We now compare the performance of individuals and groups under different protocols, accounting for both decision quality and information costs.

We normalize the payoff for a correct guess to 1, the cost to 0.2, and divide the time waited in seconds by 60. Utilizing the posterior and time of the pivotal vote, we calculate the following performance measure, corresponding to average welfare

$$\lambda_{i,g}^{benchmark} = p_{i,g} - 0.2 \cdot t_{i,g},$$

where i represents a treatment, and g represents a particular group in a particular round within the treatment. The average performance of groups voting under unanimity exceeds the performance of individuals and groups using majority, but the differences are not statistically pronounced.[18]

The performance measure assessed above necessarily inherits the randomness induced by the particular information processes participants face. For any threshold posterior p, the expected stopping time is $\mathbb{E}[t|p] = \frac{(2p-1)\log\left(\frac{p}{1-p}\right)}{\mu}$. Thus, we define the expected performance as[19]

$$\lambda_{i,g}^{expected} = p_{i,g} - 0.2 \cdot \mathbb{E}[t|p_{i,g}].$$

The reduction in noise brings forth significant differences. Performance in groups using unanimity is higher than performance generated by individuals ($p < 0.05$), and by groups using majority ($p < 0.01$).[20]

It is also possible to compare performance in the three treatments with the optimal performance under static information acquisition (classical hypothesis testing). Given our parameter

[18] Performance in our individual, majority, and unanimity treatments are 0.655(0.003), 0.650(0.005), and 0.662(0.006), respectively, with numbers in parentheses representing standard errors.

[19] Here, we effectively assume time-independent thresholds. This approximation simplifies assessments dramatically and yields results that are in line with those from our alternative performance measures.

[20] Expected performance in our individual, majority, and unanimity treatments are 0.651(0.003), 0.648(0.003), and 0.660(0.002), respectively, with numbers in parentheses representing standard errors.

values, when information collection durations are chosen at the outset, the optimal duration is 29.4 seconds, resulting in an expected posterior of 0.72, and an expected performance of 0.622. This performance is statistically significantly worse than all of our dynamic treatments. Of course, optimal performance under static collection is an upper bound on performance by real individuals. Therefore, the performance advantage of our dynamic treatments would be even larger when compared to any possible treatment under static information collection.

6 Explaining Hasty Majority Decisions

Why are decisions under majority so hasty, whereas participants' behavior under unanimity is indistinguishable from behavior in the individual treatment? This behavior seems incompatible with several prominent behavioral channels that have been emphasized in the context of group interactions. For example, altruism, or other-regarding preferences (Cooper and Kagel, 2016) would operate similarly under majority and unanimity and would not generate the differences we observe. Similarly, *diffusion of responsibility* (Darley and Latané, 1968), suggesting individuals' prefer not to affect the group's choice, would push the second and third voters in the unanimity treatment to vote faster in order to avoid being pivotal, in contrast with our data.

We propose two complementary explanations for hasty majority decisions. First, faster decisions can be an optimal response to sub-optimal decisions by other voters. As we document in Section 3, participants' behavior is heterogeneous. This heterogeneity does not affect optimal behavior as long as other agents use stationary and symmetric thresholds (treating the two states symmetrically, as our participants do).[21] However, Section 4.1 shows that many participants display declining thresholds. It turns out that a rational participant who expects her group members to have declining thresholds may be hastier under majority rule, while the theoretical optimal stationary threshold remains a best response under unanimity. We elaborate on this effect in the Online Appendix.

The second possible explanation for hasty majority decisions relates to recent work suggesting that individuals desire to influence outcomes, a *demand for agency*. See, for instance, Fehr et al. (2013), Bartling et al. (2014), and Pikulina and Tergiman (2020). This force does not affect individuals acting on their own and is not likely to influence groups using unanimity: in both cases,

[21] Across rounds and treatments, participants vote in favor of states A and B with frequencies of 49.5% and 50.5%, respectively.

Table 3: Difference in Posterior: Second vs First Voter

	$(p_2 - p_1)$
Constant	0.193***
	(0.0176)
d_M	-0.154***
	(0.0338)
d_U	-0.00695
	(0.0205)
p_1	-0.607***
	(0.0639)
$p_1 \times d_M$	0.164***
	(0.0548)
$p_1 \times d_U$	0.00585
	(0.0288)
N	330960

Standard errors in parentheses
Process-level clustering
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: We run a regression of the difference between the posterior with which the second and first vote is cast, on treatment indicators (d_M for majority and d_U for unanimity), the absolute difference between the first vote's posterior and 0.5 (p_1), interactions of the first vote's posterior and treatments ($p_1 \times d_M$ and $p_1 \times d_U$), and a constant.

a decision can only be made after each participant has cast a vote. In contrast, under majority rule, the group decision is made by the majority of participants who are quickest to vote. Thus, a desire to affect the group decision may lead to hastier decisions.[22]

To evaluate the plausibility of a demand for agency, we inspect remaining voters' responses to the first vote being cast. Under majority rule, a demand for agency would introduce a race between the remaining two group members and reduce the posterior at which the second vote is cast. Table 3 displays the results of a regression in which, within each combination of treatment, group, and round, we calculate the difference between the posterior at which the second vote and the posterior at which the first vote was cast.

The variables d_M and d_U are dummy variables corresponding to the majority and unanimity treatments, respectively. The variable p_1 stands for the first stopping posterior, normalized by subtracting 0.5. Thus, the intercept corresponds to the *additional* accuracy required by the second voter when the first voter casts a vote with a posterior of 0.5. The variables $p_1 \times d_M$ and $p_1 \times d_U$ correspond to the interactions between p_1 and the corresponding treatment dummies, allowing for different slopes across treatments.[23]

[22] An alternative *demand for pivotality*, whereby individuals wish to be decisive, would lead first voters in the majority treatment and second voters in the unanimity treatment to delay their votes, which is inconsistent with our data.

[23] To estimate the parameters, we rely on choices of individuals across sessions. Therefore, we cannot cluster errors at the session level. We take a conservative approach and cluster errors at the process level. For our control, the individual treatment, we rely on simulated groups based on the procedure described in Section 4.2.

As can be seen, d_M and $p_1 \times d_M$ are both statistically significant at the 1% level: second voters under majority are hastier than simulated second voters based on the individual treatment. This is not the case for group members operating under unanimity.

Given the observed responsiveness of second voters to first voters' choices under majority, there is a strategic reason for first voters to expedite their choices as well. Relatively lenient agents, who are likely to be first voters, can manipulate the pivotal threshold posteriors to be closer to their desired thresholds. With a demand for agency, by expediting her choice, a more lenient member induces a hastier second vote. Ideally, the first voter would tailor her stopping posterior so that the pivotal vote would occur at precisely her desired threshold. This is consistent with our observations: the right panel of Figure 2 indicates that the distribution of posteriors when the second, pivotal votes are cast under majority closely approximates the distribution of posteriors when the first, most lenient votes occur under both the unanimity and the (simulated) individual treatments.

7 Conclusions

This paper reports results from a set of experimental treatments testing sequential sampling, in individuals and groups. In our experiments, groups operating under unanimity deliver the best outcomes. Contrasting theoretical predictions, sequential sampling yields time-decreasing thresholds. Furthermore, groups behave differently than individuals, with majority rule yielding the quickest decisions.

Our experimental framework and our results point to several possible future directions of inquiry. In our study of individuals, it would be interesting to consider behavior under a richer set of parameters. In our study of groups, we have focused on a baseline case in which the model predicts no group effects. This is intended as an initial benchmark on which to build a richer understanding of heterogeneous committees, as in the model studied in Chan et al. (2018). Similar experiments could be designed with group members experiencing heterogeneous preferences over alternatives and heterogeneous information costs. It would also be interesting to vary other features of groups: their size, the monitoring available to group members, etc. Finally, it would be useful to design additional experiments to inspect the generality of the non-stationarity we observe and investigate the demand for agency directly.

References

Baldassi, C., Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., and Pirazzini, M. (2020). A behavioral characterization of the drift diffusion model and its multialternative extension for choice under time pressure. *Management Science*, 66(11):4921–5484.

Bartling, B., Fehr, E., and Herz, H. (2014). The intrinsic value of decision rights. *Econometrica*, 82(6):2005–2039.

Becker, G. M. (1958). Sequential decision making: Wald's model and estimates of parameters. *Journal of Experimental Psychology*, 55(6):628.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186.

Bhattacharya, S., Duffy, J., and Kim, S. (2017). Voting with endogenous information acquisition: Experimental evidence. *Games and Economic Behavior*, 102:316–338.

Brown, M., Flinn, C. J., and Schotter, A. (2011). Real-time search in the laboratory and the market. *American Economic Review*, 101(2):948–74.

Caplin, A., Dean, M., and Martin, D. (2011). Search and satisficing. *American Economic Review*, 101(7):2899–2922.

Chan, J., Lizzeri, A., Suen, W., and Yariv, L. (2018). Deliberating collective decisions. *Review of Economic Studies*, 85(2):929–963.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, S. and Heese, C. (2021). Fishing for good news: Motivated information acquisition. *mimeo*.

Cooper, D. J. and Kagel, J. H. (2016). Other-regarding preferences. *The Handbook of Experimental Economics*, 2.

Darley, J. M. and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4p1):377.

Dominitz, J. and Manski, C. F. (2017). More data or better data? a statistical decision problem. *The Review of Economic Studies*, 84(4):1583–1605.

Dvoretzky, A., Kiefer, J., Wolfowitz, J., et al. (1953). Sequential decision problems for processes with continuous time parameter. testing hypotheses. *The Annals of Mathematical Statistics*, 24(2):254–264.

Edwards, W. and Slovic, P. (1965). Seeking information to reduce the risk of decisions. *The American Journal of Psychology*, 78(2):188–197.

Elbittar, A., Gomberg, A., Martinelli, C., and Palfrey, T. R. (2020). Ignorance and bias in collective decisions. *Journal of Economic Behavior & Organization*, 174:332–359.

Fehr, E., Herz, H., and Wilkening, T. (2013). The lure of authority: Motivation and incentive effects of power. *American Economic Review*, 103(4):1325–59.

Freer, M., Martinelli, C., and Wang, S. (2020). Collective experimentation: A laboratory study. *Journal of Economic Behavior & Organization*, 175:365–379.

Fried, L. S. and Peterson, C. R. (1969). Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology*, 80(3p1):525.

Fudenberg, D., Strack, P., and Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–3684.

Gabaix, X., Laibson, D., Moloche, G., and Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4):1043–1068.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Ginzburg, B. and Guerra, J.-A. (2019). When collective ignorance is bliss: Theory and experiment on voting for learning. *Journal of Public Economics*, 169:52–64.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112(2):631–645.

Green, P. E., Halbert, M. H., and Minas, J. S. (1964). An experiment in information buying. *Journal of Advertising Research*, 4(3):17–23.

Grosser, J. and Seebauer, M. (2016). The curse of uninformed voting: An experimental study. *Games and Economic Behavior*, 97:205–226.

Henry, E. and Ottaviani, M. (2019). Research and the approval process: The organization of persuasion. *American Economic Review*, 109(3):911–55.

Luce, R. D. et al. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford University Press on Demand.

McClellan, A. (2021). Experimentation and approval mechanisms. *mimeo.*

Pikulina, E. S. and Tergiman, C. (2020). Preferences for power. *Journal of Public Economics*, 185:104173.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922.

Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2):333.

Strack, P. and Viefers, P. (2021). Too proud to stop: Regret in dynamic decisions. *Journal of the European Economic Association*, 19(1):165–199.

Strulovici, B. (2010). Learning while voting: Determinants of collective experimentation. *Econometrica*, 78(3):933–971.

Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics*, 12(1):16–32.

Wald, A. (1945). Sequential method of sampling for deciding between two courses of action. *Journal of the American Statistical Association*, 40(231):277–306.

Wald, A. (1947). *Sequential Analysis*. New York.