Experimental Economics (2025), 1–26
doi:10.1017/eec.2024.9

ORIGINAL PAPER

Learning to detect change: an experimental investigation

Ye Li[1], Cade Massey[2] and George Wu[3]

[1]University of California, School of Business, Riverside, CA USA
[2]University of Pennsylvania, Wharton School of Management, Philadelphia, PA
[3]University of Chicago, Booth School of Business, Chicago, IL
Corresponding author: Ye Li; Email: ye.li@ucr.edu

(Received 31 August 2023; revised 1 November 2024; accepted 1 November 2024)

Abstract

People, across a wide range of personal and professional domains, need to accurately detect whether the state of the world has changed. Previous research has documented a systematic pattern of over- and under-reaction to signals of change due to *system neglect*, the tendency to overweight the signals and underweight the system producing the signals. We investigate whether experience, and hence the potential to learn, improves people's ability to detect change. Participants in our study made probabilistic judgments across 20 trials, each consisting of 10 periods, all in a single system that crossed three levels of diagnosticity (a measure of the informativeness of the signal) with four levels of transition probability (a measure of the stability of the environment). We found that the system-neglect pattern was only modestly attenuated by experience. Although average performance did not increase with experience overall, the degree of learning varied substantially across the 12 systems we investigated, with participants showing significant improvement in some high diagnosticity conditions and none in others. We examine this variation in learning through the lens of a simple linear adjustment heuristic, which we term the "δ-ϵ" model. We show that some systems produce consistent feedback in the sense that the best δ and ϵ responses for one trial also do well on other trials. We show that learning is related to the consistency of feedback, as well as a participant's "scope for learning" how close their initial judgments are to optimal behavior.

Keywords: Bayesian updating; Change-point detection; learning; over-reaction; regime shift; under-reaction

JEL Codes: C91; D84; D91

1. Introduction

The need to detect change accurately is a common problem for people in a wide range of domains, from business and politics to social relations and sports. In the academic literature, the canonical example is monitoring quality levels in a manufacturing process (Shewhart, 1939, Girshick & Rubin, 1952, Deming, 1975), but recent research has broadened this paradigm beyond the field of operations research. Researchers in finance have invoked regime-shifts to explain documented patterns of under- and overreaction in asset pricing (Barberis et al., 1998, Brav & Heaton, 2002, Gennaioli et al., 2015), and economists have used change-point models to describe the challenges central bankers face in setting interest rates (Ball, 1995, Blinder & Morgan, 2005, Hamilton, 2016, Brunnermeier et al., 2021). Even further afield, change detection is important to retailers assessing changes in consumer taste (Fader & Lattin, 1993), corporate strategists monitoring technological trends in their marketplace (Grove, 1999), politicians tracking voter sentiment (Bowler & Donovan, 1994), and individuals attending to their health (Steineck et al., 2002) or their romantic partner's commitment (Sprecher, 1999). Finally and most recently, epidemiologists, policy makers, and the general population spent much of the COVID-19 pandemic figuring out whether a surge had peaked or a new COVID variant had taken hold (Atkeson et al., 2021, Yang et al., 2021). In all of these cases, "the states of the process are not directly observable but become gradually known with the sequential acquisition of fallible information over time" (Rapoport et al., 1979). Indeed, the need to detect change accurately is ubiquitous, and it is therefore critical to understand the behavioral patterns involved in doing so.

These examples highlight the challenge of successfully identifying change: One must infer the true state or "regime" from unreliable signals, while balancing the costs of under-reacting (failing to realize change has occurred) against the costs of over-reacting (believing change has occurred when in fact it has not). For example, an investor needs to recognize when cyclical financial markets change from a "bear" to a "bull" market. Economic indicators are, at best, imprecise signals, with informativeness varying across indicators and over time. Under-reacting to signals of change means foregoing the chance to buy stocks at their lowest prices (or failing to sell a stock when it has peaked), whereas over-reacting entails acquiring shares that are still declining (or selling a stock too early when it would otherwise continue to rise).

A number of psychologists have investigated how successfully individuals navigate various experimental instantiations of change point detection tasks (*i.e.*, Robinson, 1964, Chinnis & Peterson, 1968, 1970, Theios et al., 1971, Barry & Pitz, 1979, Rapoport et al., 1979, Estes, 1984, Brown & Steyvers, 2009). A theme in this literature is that individuals respond to environmental conditions, but only partially. As Chinnis & Peterson (1968) stated: "subjects, while sensitive to the difference in diagnostic value of the data in the two conditions, were not adequately sensitive" (p. 625). Massey & Wu (2005a) expanded on this theme, proposing the *system-neglect hypothesis*: People react primarily to the signals they observe and secondarily to the system that produced the signal. In their experiments, participants were exposed to signals generated by a number of different systems in which diagnosticity (*i.e.*, the precision of the signal) and transition probability (*i.e.*, the stability of the system) were varied. In our investor example, diagnosticity corresponds to the informativeness of the market indicators and transition probability corresponds to the historical rate of market vacillation. A diagnostic (vs. undiagnostic) signal might be an analyst report from a well-respected and historically successful (vs. unknown) securities analyst, or alternatively many disparate signals that largely point in the same direction (vs. a small number of disparate signals that are somewhat contradictory).[1] A stable (vs. unstable) system might be a more developed market such as the U.S. (vs. an emerging market such as Brazil).

Massey and Wu's laboratory studies revealed a behavioral pattern consistent with system neglect: Under-reaction was most common in unstable systems with precise signals, whereas over-reaction was most prevalent in stable systems with noisy signals. Kremer et al. (2011) replicated and extended their work, finding evidence of system neglect in a time-series environment with continuous change and continuous signals. More recently, Seifert et al. 2023 examined pricing decisions with impending regime shifts. They found a pattern similar to Massey and Wu in a rich setting in which there were buyers and sellers, a regime shift was either desirable or undesirable, and outcomes were either gains or losses (see also, Guha et al., 2024).[2]

A common critique of behavioral decision research is that participants engage in relatively novel tasks in unfamiliar environments, without sufficient opportunity to learn (*i.e.*, Coursey et al., 1987, Hertwig & Ortmann, 2001, List, 2003, Plott & Zeiler, 2005, Erev & Haruvy, 2015). In Massey & Wu (2005a) (hereafter MW), for example, the diagnosticity and transition probability of the system changed after each of the 18 trials. On the one hand, this design enhanced

[1] Diagnosticity might also reflect the aggregation of expert opinions, with diagnosticity reflecting the quality of the cues used by the experts, as well as the inter-cue correlation. Budescu & Yu (2007) documented a pattern consistent with system neglect in their empirical investigation of this non-probabilistic environment.

[2] See also Wang et al. (2024) for a replication of Massey & Wu (2005a) and an identification of some neural correlates of system neglect.

the salience of the system variables, increasing the likelihood that participants would give them sufficient attention. On the other hand, this continuously changing design may have hindered participants' ability to appropriately adjust to these dimensions by minimizing their opportunity to learn about a particular system and therefore raises questions about the robustness of the system-neglect hypothesis. Does the pattern of over- and under-reaction observed in MW persist in the face of the opportunity to learn? Or is the system-neglect phenomenon only observed in those inexperienced with the task and therefore less applicable to real-world settings where people have sufficient opportunities to learn? The present paper aims to fill these gaps in our understanding.

Although there is little doubt that experience can lead to learning, it is also well known that it does not always, especially for tasks involving probabilistic judgment (Brehmer, 1980). Experimental studies in psychology and economics have found that experience *can* improve performance in probabilistic tasks such as conditional probability estimation (Martin & Gettys, 1969, Donnell & Du Charme, 1975), the Monty Hall Problem (Friedman, 1998), and the Newsvendor Problem (Bolton & Katok, 2008). However, learning is often limited and depends on the nature of the feedback (Martin & Gettys, 1969, Hogarth et al., 1991, Hogarth, 2001). In particular, Hogarth et al. (1991) found that the "exactingness" (*i.e.*, the severity of penalties imposed for errors) of feedback had an inverted-U-shaped relationship with learning: while some penalty for errors helps, overly exacting feedback reduces learning. Importantly, this prior work has focused on static tasks in which the task parameters do not change over time; learning in dynamic tasks may be more limited since feedback across trials may not be relevant for future decisions (Schweitzer & Cachon, 2000). While we expect feedback to be important for learning to detect change, what system characteristics give rise to consistent feedback with the right degree of exactingness?

The rest of the paper is organized as follows. We begin by describing the statistical process used in our experiment and reviewing the system-neglect hypothesis. Next, we present the experimental design, in which participants made probability judgments across 20 trials, each consisting of 10 periods, all in a single system that crossed three levels of diagnosticity (a measure of the informativeness of the signal) with four levels of transition probability (a measure of the stability of the environment). We discuss our results and examine how learning varies across conditions. We focus on two measures, earnings and reactions to signals of change. We then present a linear adjustment heuristic, which we term the δ-ϵ model, that provides good fits to both the normative standard of Bayesian updating, as well as the probabilities elicited from our participants. We use the δ-ϵ model to show that environments vary considerably in how exacting they are to deviations from the optimal δ-ϵ strategy, and in the consistency of feedback they provide how the earnings-maximizing δ and ϵ for one trial (a "local maxima") performs on other trials. We show that a substantial portion of the differences in learning across environments can be explained by feedback consistency for that environment, as well as by "scope for learning" how close or far participants' initial reactions (in the sense of δ and ϵ) are to optimal reactions. We conclude by discussing open questions and future directions.

2. Background and Theory

In this section, we introduce the design of our experiment and review the system-neglect hypothesis predictions for detecting changes in this statistical process.

2.1. Terminology

We begin by introducing key terms used throughout the paper. First, the *system* is the random process that generates binary *signals* (red or blue balls) in each of the 10 *periods* that make up a *trial*. Systems are dynamic in the sense that they can generate signals using two different sets of probability distributions, each of which we call a *regime*. A system is characterized by two system parameters:

diagnosticity, or the informativeness of the signals it generates, and *transition probability*, or the likelihood of the system switching to the second regime.

2.2. Experimental Paradigm

Our experimental paradigm largely mirrors that of MW. Each trial t consists of 10 periods, with the system beginning in the red regime. There is, however, a transition probability q of switching to the blue regime before any period i (including the first period before any signal is drawn). If the system switches to the blue regime, it does not switch back. That is, the blue regime is an absorbing state.[3] The regime changes at $\tau_t = 1, \ldots, 11$, where $\tau_t = 1$ indicates that the regime changes before the first signal is received and $\tau_t = 11$ indicates that the regime did not shift during that trial.

The system generates either a red or blue signal in each period. A red signal is generated by the red regime with probability $p_R > .5$ and by the blue regime with probability $p_B < .5$. Put differently, a red signal is more suggestive of a red regime, and a blue signal is more suggestive of a blue regime. The probabilities were symmetric in our experiment (*i.e.*, $p_R = 1 - p_B$), so p_R/p_B is a measure of the diagnosticity (d) of the signal, with larger diagnosticities corresponding to more precise and informative signals. Participants were given each of the relevant system parameters and told that their task was to guess which regime generated that period's signal. More specifically, they estimated the probability that the system had switched to the blue regime. Importantly, at the end of each trial, participants received feedback about the true regime that governed each period of that trial.

Optimal responses to the task required application of Bayes' Rule. Let $B_i = 1$ ($B_i = 0$) indicate that the stochastic process is in the blue (red) regime in period i, and let $b_i = 1$ ($b_i = 0$) indicate that a blue (red) signal is observed in that period. If $H_i = (b_1, \ldots, b_i)$ denotes the history of signals through period i, the Bayesian posterior odds of a change to the blue state after observing history H_i is:

$$\frac{p_i^b}{1 - p_i^b} = \left(\frac{1 - (1-q)^i}{(1-q)^i}\right) \sum_{j=1}^{i} \frac{q(1-q)^{j-1}}{1-(1-q)^i} d^{\left[i+1-j-\left(2\sum_{k=j}^{i} b_k\right)\right]}, \tag{1}$$

where $p_i^b = \Pr(B_i | H_i)$ denotes the Bayesian posterior probability that the process has switched to the blue regime by period i. The derivation for Eqn. (1) is found in Massey & Wu (2005b). Note also that the right hand side of Eqn. (1) factors out $(1 - (1-q)^i))/(1-q)^i$, the "base rate" odds of a change to the blue regime in the absence of a signal.

Our experimental setting allowed us to compare individual judgments against the normative standard of Bayesian updating, as we provided participants with all the information necessary to calculate Bayesian responses and hence provide optimal judgments. Therefore, our experiment was designed to test the system-neglect hypothesis by investigating whether individuals update probability judgments as required by Bayesian updating and whether their ability to do so improves with experience.

2.3. The system-neglect hypothesis

The system-neglect hypothesis posits that people are more sensitive to signals than to system variables. This hypothesis extends work by Griffin & Tversky (1992), who proposed that people are disproportionately influenced by the strength of evidence (*i.e.*, the effusiveness of a letter of recommendation) at the expense of its weight (*i.e.*, the credibility of the letter writer) (see also, Benjamin,

[3] Although an absorbing state is a simplification, this is a standard assumption in this literature see, however, Brown & Steyvers, 2009. An absorbing state process is also a reasonable way to model many of the low-frequency real-world examples we discuss above. To the investor who hopes to cash out before a downturn, it does not matter if a bear market would eventually return to its bull status. For the relevant decision time frame, the bear market might as well be an absorbing state.

2019). This relative attention to strength over weight determines a person's confidence, leading to a pattern of over-confidence when strength is high and weight is low, and under-confidence when strength is low but weight is high. For example, Griffin and Tversky, 1992 investigated a static analog to our task in which participants judge whether a coin is biased heads or biased tails based on a sample of h heads out of n tosses. They found that judgments are more influenced by the sample proportion (h/n), the strength of evidence, than the sample size (n), the weight of evidence. In another study, sample proportion again constituted the strength of evidence, with base rate serving as the weight of evidence in this instance.

In the context of our dynamic statistical process, we take the signal (*i.e.*, the sequence of red and blue signals) to be the strength and the system parameters (*i.e.,* the transition probability, q, and the diagnosticity, d) to be the weight. The critical implication of the system-neglect hypothesis is that individuals are more likely to over-react to signals of change in stable systems with noisy signals, and are more likely to under-react in unstable systems with precise signals. However, note that system neglect makes a relative prediction and is silent about overall levels of reaction; as such, it is consistent with patterns of only under-reaction or only over-reaction.

To provide a concrete example, consider four systems crossing two levels of diagnosticity, $d = 1.5$ and $d = 9$, with two transition probabilities, $q = .05$ and $q = .20$. Suppose that signals in the first two periods are both blue, *i.e.*, $H_2 = (1, 1)$. The Bayesian posterior probabilities of a change to the blue regime are $\Pr(B_2|H_2) = .17$ when $d = 1.5$ and $q = .05$ (*i.e.*, a noisy and stable system), but $\Pr(B_2|H_2) = .92$ when $d = 9$ and $q = .20$ (*i.e.*, a precise and unstable system). If individuals give approximately the same response across all four conditions (for example, with a posterior probability of .60), they will over-react when $d = 1.5$ and $q = .05$ and under-react when $d = 9$ and $q = .20$. Of course, we do not expect participants to ignore the system parameters entirely. However, the system-neglect hypothesis requires that people attend too little to diagnosticity and transition probability and too much to the signals.

3. Learning Experiment

3.1. Participants, Stimuli, and Conditions

We recruited 240 students at a private Midwestern university as participants for a task advertised as a "probability estimation task." Each participant was randomly assigned to one of the 12 experimental conditions, constructed by crossing three diagnosticity levels ($d = 1.5$, 3, and 9) with four transition probability levels ($q = .02$, .05, .10, and .20). These 12 systems were the same ones used in MW. The most important deviation from MW was switching to a between-participants design: Each participant only saw one set of system variables for all 20 trials.

Although we pre-generated the 20 random trials of 10 periods each using each condition's system variables, participants received those 20 trials in randomized order. The actual series for each trial can be found in the Online Supplemental Materials (see A.1).[4]

3.2. Compensation

We compensated participants according to a quadratic scoring rule that paid a maximum of \$0.08 (*i.e.*, if a participant indicated a 100% probability that the system was in the blue regime and it in fact was) and a minimum of -\$0.08 (*i.e.*, if a participant indicated a 100% probability that the system was in the blue regime but it was in fact still in the red regime). A quadratic scoring rule theoretically

[4] The computer program, data, and statistical code have been posted at: https://osf.io/5vy9m/?view_only=2daf13df4b124901a16f2da26b353505.

elicits true beliefs for a risk-neutral participant (, but see). Although it was possible to lose money overall, doing so was extremely unlikely.[5] There was no fixed show-up fee.

3.3. Method

The experiment was conducted using a specially-designed Visual Basic program (see Section B in the Online Supplemental Materials for experiment screenshots). The program began by introducing the statistical process used in the experiment, explaining the system variables (p_R, p_B and q), how the computer would pick balls (*i.e.*, the signals) from one of two bins (*i.e.*, the regimes), and how the bin may switch. The program showed a schematic diagram of bin switching and then displayed four demonstration trials, each consisting of ten sequential draws. For these demonstration trials only, participants saw the actual sequence of bins that generated each signal, and therefore, if and when the process shifted from the red to the blue bin.

Participants were then told that, after seeing each ball, their task was to estimate the probability that the system had switched to the blue bin (*i.e.*, the probability that the regime had shifted) by entering any number between 0 and 100. The computer then gave a detailed explanation of the incentive procedure including payment curves as a function of estimated probability and whether the bin had actually switched to the blue bin by that period. Participants completed two unpaid trials to better understand the interface and the incentive structure. After each trial, they were informed how much money they would have made or lost on that particular trial. At the end of each paid trial, participants received feedback about which bin generated each ball, earnings for that trial, and cumulative earnings in the experiment. Finally, participants completed 20 trials for actual pay.

4. Results

In this section, we summarize basic results for performance and learning. We look at earnings as a measure of performance. We report on Mean Absolute Difference (MAD) between empirical and Bayesian judgments in the Online Supplemental Materials, Section C.2. To test the system-neglect hypothesis, we then consider measures of reaction to indications of change as in MW.

In Section 5, we analyze the results through the lens of a heuristic model of δ-ϵ adjustment. This relatively simple "linear adjustment" model can mimic both Bayesian probabilities, as well as participants' judgments. We also use this model to characterize the consistency of feedback within a system environment and show how this characterization relates to learning (or lack thereof) across conditions.

4.1. Earnings

Recall that we paid participants based on the square of the difference between their subjective probability of having changed to the blue regime and the actual regime in that period (1 if blue, 0 if red). The mean absolute deviation (MAD) between their predictions and the actual regime was 0.150 ($median = 0.073$, $sd = 0.192$), generating mean total earnings of $11.05 over our 240 participants (range of $2.03 to $15.19 across participants) We define Bayesian earnings to be the earnings that would accrue if a participant's probabilities were Bayesian as in Eqn. (1). Bayesian earnings were $12.95, while a participant who gave a probability of .5 for all 200 signals would have earned $8.00. Participants under-performed Bayesian earnings by a mean of $1.91 (range of $-0.27 to $12.34 across participants), earning 14.7% less than a Bayesian agent would earn. The rightmost column of Table 1 presents the mean empirical and Bayesian earnings overall for each condition.

[5] Massey & Wu (2005a) also used a quadratic scoring scheme, although participants in that study could make or lose as much as 10 cents in each of the 10 periods for 18 trials.

Table 1 Mean empirical and Bayesian earnings, and the difference between them, by condition and quintile (set of 4 trials). Also shown are linear regression coefficients of earnings as a function of trial, accounting for participant-level random intercepts and slopes, and the percentage of participants with improving earnings. Data by quintile and trial are re-scaled to facilitate comparison with overall earnings * $p < .05$

	Condition												
	$d = 1.5$				$d = 3$				$d = 9$				
	$q = .02$	$q = .05$	$q = .10$	$q = .20$	$q = .02$	$q = .05$	$q = .10$	$q = .20$	$q = .02$	$q = .05$	$q = .10$	$q = .20$	Overall
Empirical Earnings													
Mean	14.35	9.69	7.88	6.59	12.36	10.99	10.55	8.84	14.15	12.13	12.08	12.96	11.05
Standard Deviation	0.88	1.39	1.61	1.71	0.84	1.26	1.32	2.50	0.57	1.64	2.69	1.75	2.82
Bayesian Earnings													
Earnings	14.92	11.47	10.21	9.85	13.49	12.73	12.44	12.20	15.07	14.05	14.74	14.28	12.95
Empirical/Bayesian	96.2%	84.5%	77.2%	66.9%	91.6%	86.3%	84.8%	72.5%	93.9%	86.3%	82.0%	90.8%	84.4%
Empirical Earnings by Quintile (scaled by 5)													
Quintile 1	14.19	8.85	8.07	6.79	13.01	10.98	9.87	7.82	12.91	11.52	11.87	13.18	10.73
Quintile 2	14.14	10.14	8.57	6.55	11.23	11.69	10.58	8.96	14.32	11.80	11.58	12.94	11.02
Quintile 3	14.24	9.79	7.68	6.61	12.36	10.41	10.90	9.17	14.31	12.60	11.47	13.14	11.09
Quintile 4	13.92	10.45	6.32	5.72	13.09	11.57	10.61	9.74	15.01	12.62	12.16	13.27	11.16
Quintile 5	14.78	9.51	8.24	7.42	12.10	10.57	11.38	8.71	14.22	12.36	13.52	12.83	11.33
Bayesian Earnings by Quintile (scaled by 5)													
Quintile 1	14.84	10.85	10.50	8.82	13.33	12.49	11.54	11.45	15.04	14.34	14.78	14.61	12.70
Quintile 2	14.89	11.59	10.19	9.54	13.23	12.86	12.14	12.28	14.98	13.48	14.81	13.92	12.80
Quintile 3	14.98	11.66	10.15	10.21	13.02	12.54	12.61	12.82	15.04	14.21	14.76	14.45	13.07
Quintile 4	14.66	11.52	10.12	9.80	14.21	12.91	12.82	12.28	15.41	14.10	14.47	14.38	13.02
Quintile 5	15.06	11.97	9.98	10.74	13.55	13.06	13.18	11.87	14.91	13.95	14.87	14.07	13.12

(*Continued*)

Table 1 (*Continued.*)

	Condition												
	$d = 1.5$				$d = 3$				$d = 9$				
	$q = .02$	$q = .05$	$q = .10$	$q = .20$	$q = .02$	$q = .05$	$q = .10$	$q = .20$	$q = .02$	$q = .05$	$q = .10$	$q = .20$	Overall
Relative Earnings (Empirical-Bayesian) by Quintile (scaled by 5)													
Quintile 1	-0.65	-2.00	-2.43	-2.03	-0.33	-1.52	-1.67	-3.63	-2.13	-2.82	-2.91	-1.43	-1.97
Quintile 2	-0.75	-1.45	-1.62	-2.99	-1.99	-1.17	-1.57	-3.32	-0.67	-1.67	-3.23	-0.98	-1.78
Quintile 3	-0.74	-1.87	-2.47	-3.61	-0.67	-2.13	-1.71	-3.65	-0.73	-1.61	-3.29	-1.31	-1.98
Quintile 4	-0.74	-1.07	-3.80	-4.07	-1.12	-1.35	-2.21	-2.54	-0.39	-1.48	-2.31	-1.10	-1.87
Quintile 5	-0.28	-2.46	-1.73	-3.31	-1.45	-2.49	-1.80	-3.17	-0.69	-1.58	-1.35	-1.24	-1.79
Change in relative earnings by trial (scaled by 200)													
Regression coefficient	0.19	-0.09	-0.22	-0.88	-0.13	-0.31	-0.30	0.66	0.62	0.68	1.01	0.20	0.12
Standard error	0.23	0.44	0.51	0.54	0.36	0.38	0.37	0.42	0.26*	0.34*	0.45*	0.22	0.11
Change in relative earnings by trial (scaled by 200)													
% Positive Learning	45%	55%	50%	30%	45%	50%	45%	65%	65%	85%	70%	50%	55%

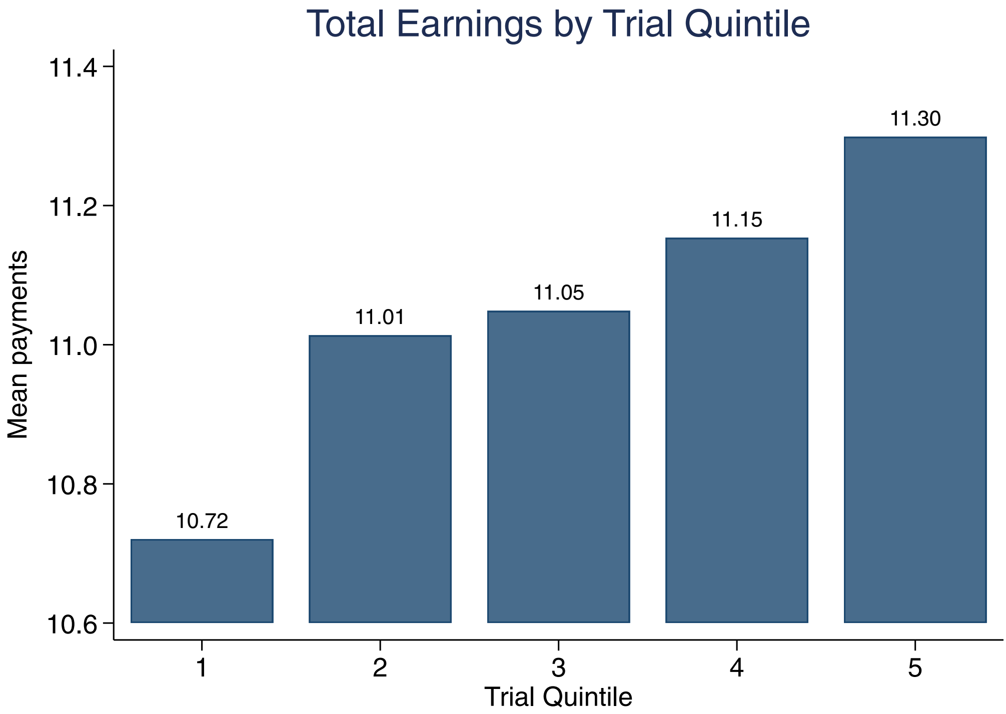
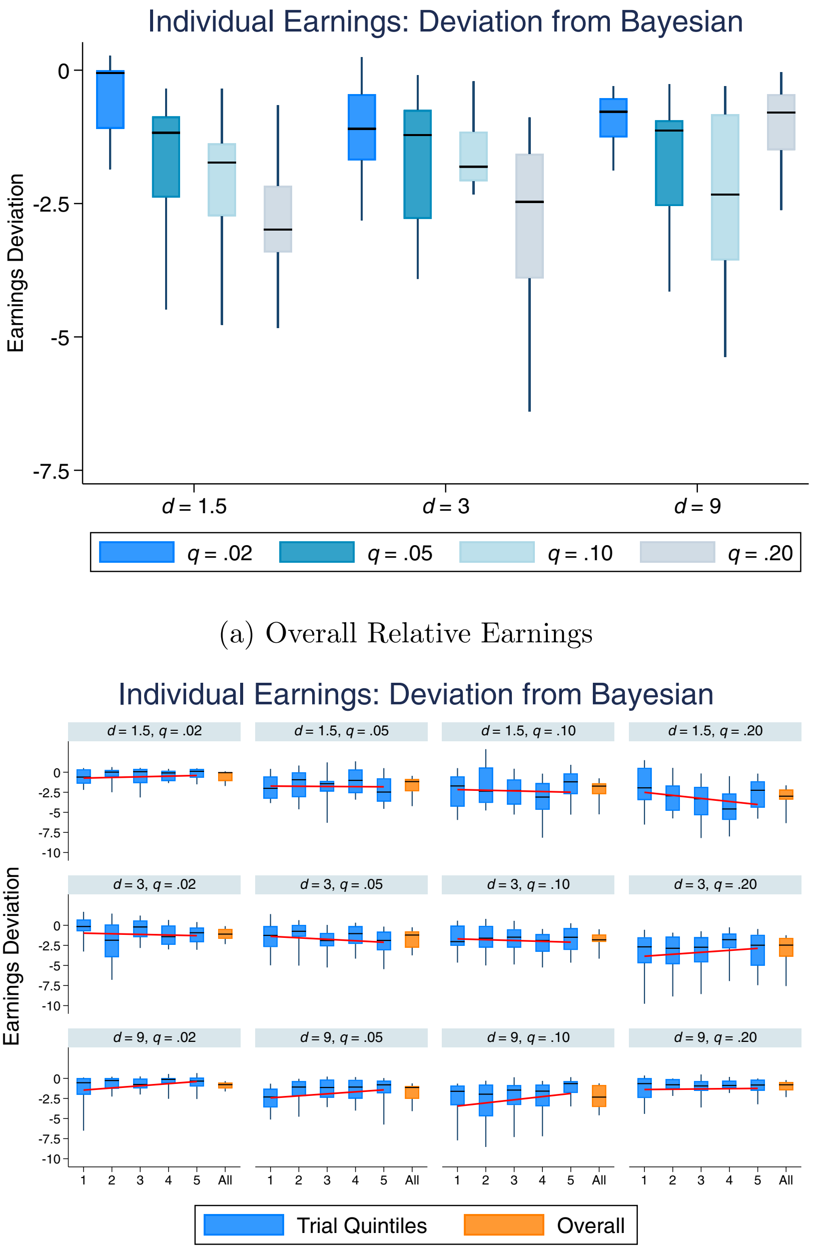


Fig 1. Average total earnings by trial quintile, where Quintile 1 consists of the first four trials experienced by a participant, etc. Quintile earnings are normalized (multiplied by five) so that they are comparable to the total earnings over all 20 trials

To investigate whether earnings increased over the course of 20 trials, we examined how earnings changed by trial quintile or quintile for short, where Quintile 1 consists of the first four trials experienced by a participant, etc. (results are virtually identical if we use quarters consisting of 5 trials). Table 1 lists the earnings by quintile, multiplied by five to be comparable to the overall earnings. Figure 1 shows that earnings increase monotonically over the quintiles, with earnings in Quintile 5 significantly larger than in Quintile 1 ($t = 2.38, p = 0.018$). There was directional evidence of learning in 7 of the 12 conditions, but a significant increase in earnings only in the $d = 3, q = .10$ ($t = 3.58, p = 0.002$) condition.

Since the random series are drawn from a noisy process—we also plot *relative* earnings—the difference between Bayesian earnings and empirical earnings—in Figure 2, (a) overall and (b) broken down by quintiles. Overall, relative earnings decreases with q, with the exception of the $d = 9$ condition. In addition, earnings are close to Bayesian for the $d = 1.5, q = .02$ and $d = 9, q = .02$ conditions, and furthest from Bayesian for the $d = 1.5, q = .20$ and $d = 3, q = .20$ conditions. The plot of relative earnings by quintile again indicates that learning, if any, is modest and heterogeneous. Relative earnings increased directionally from Quintile 1 to Quintile 5 in 7 of the 12 conditions.

We conducted a statistical test of learning by running linear regressions of relative earnings on trial order for the 400 participant-trial observations per condition (20 participants per condition × 20 trials per participant), accounting for participant-level random intercepts and slopes. The coefficients on trial are found at the bottom of Table 1, with positive numbers corresponding to improving performance. Coefficients in 6 of the 12 conditions were positive and significant at the .05 level in the $d = 9, q = .02$, .05, and .10 conditions. There was no evidence for learning overall ($z = 1.13, p = 0.46$), We conducted the same analysis separately for each of the 240 participants,



(a) Overall Relative Earnings

(b) Relative Earnings by Trial Quintile

Fig 2. Relative earnings, by condition and trial quintile, with the box representing the inter-quartile range and the whiskers representing the range from 10 to 90 percentile of individuals. Relative earnings are the difference between empirical earnings and Bayesian earnings (*i.e.*, what a Bayesian agent would earn). Panel (a) shows relative earnings aggregated over the 20 trials; Panel (b) shows the same measure for each of the five quintiles (blue), as well as overall (orange)

finding positive coefficients for 131 (55%) of the 240 participants overall ($p = 0.175$, two-tailed binomial test vs. 50%). The percentages of participants who had positive regression coefficients in each condition are shown at the bottom of Table 1.

In sum, we find modest evidence for learning in earnings and significant learning only in some of the high-diagnosticity conditions. Note, however, that earnings are a potentially noisy measure of performance, since the signals that participants receive can be unrepresentative of the underlying regimes that determine their earnings. Indeed, Bayesian judgments can produce lower earnings than non-Bayesian judgments for small, unrepresentative sequences. In C.2 in the Online Supplemental Materials, we show more pronounced, but still modest, learning for the Mean Absolute Deviation (MAD) between empirical and Bayesian probabilities.

4.2. Reactions

Whereas our analyses of earnings demonstrated modest learning that varied across conditions, the system-neglect hypothesis specifies how empirical probability judgments *react* to indications of change (*i.e.*, blue signals) rather than their absolute levels. Therefore, to test for system neglect, we compared participants' reactions (*i.e.*, the change in probability judgments from period $i-1$, p_{i-1}^e, to period i, p_i^e, $\Delta p_i^e = p_i^e - p_{i-1}^e$) to the Bayesian reaction using the participant's probability judgment in the previous period as the "prior." That is, we constructed the Bayesian reaction by taking p_{i-1}^e as the "prior" and applying Bayes' Rule:

$$\frac{\bar{p}_i^b}{1-\bar{p}_i^b} = \frac{p_{i-1}^e}{1-p_{i-1}^e}\left(\frac{1}{1-q}\right)\left(\frac{p_R}{p_B}\right)^{2b_i-1} + \left(\frac{q}{1-q}\right)\left(\frac{p_R}{p_B}\right)^{2b_i-1}. \qquad (2)$$

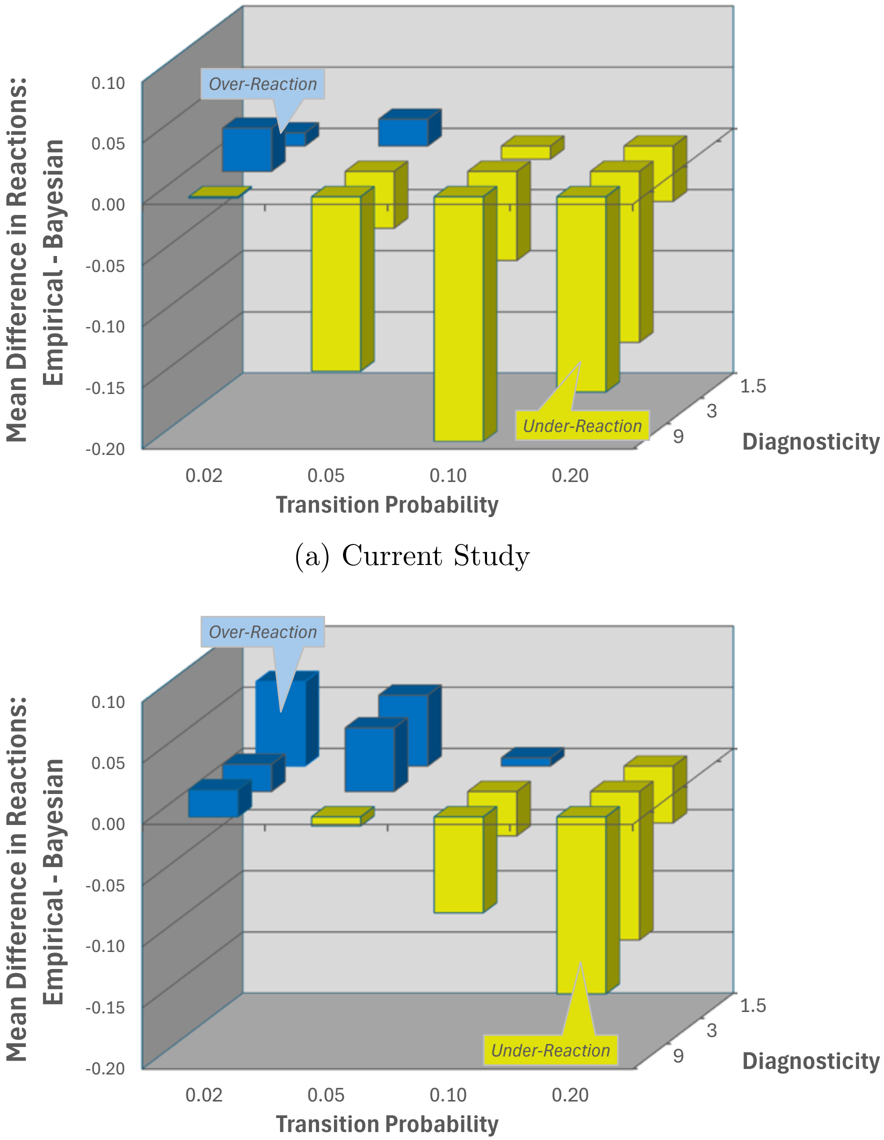
Importantly, this approach focuses only on reactions, granting participants their priors regardless of accuracy, and evaluating only how their judgments react to new information. We define errors in reaction as the difference between empirical and Bayesian reactions, $p_i^e - \bar{p}_i^b$, with under-reaction indicating an empirical reaction that is less positive than the Bayesian reaction, $p_i^e - \bar{p}_i^b < 0$, and over-reaction indicating the opposite, $p_i^e - \bar{p}_i^b > 0$.

Figure 3 depicts the mean error in reactions to blue signals by condition (red signals are indicators of "non-change" and exhibit a different gradient; see Massey & Wu 2005a). Recall that the system-neglect hypothesis predicts a greater tendency to under-react in more precise, less stable conditions, and to over-react in noisier, more stable conditions.

As predicted by the system-neglect hypothesis, and replicating MW, the greatest under-reaction occurred in the southeast-most cells ($d = 9$ with $q = .05$, $q = .10$, and $q = .20$, and $d = 3$ with $q = .20$), while the greatest over-reaction occurred in the northwest-most cells ($d = 1.5$ with $q = .02$ and $q = .05$ and $d = 3$ with $q = .02$). For 43 of the 48 pairwise comparisons between conditions, under-reaction increased monotonically with diagnosticity and transition probability ($p < .001$, two-tailed binomial test) (see Online Supplementary Materials, Table A15, for t-statistics for all 48 comparisons). For comparison, Figure 3 also plots the reactions from Massey & Wu (2005a). Note that the pattern of system neglect is somewhat more pronounced in that study compared to the current investigation.

Figure 4 plots the same errors in reactions to blue signals by quintile. Note that the degree of system neglect is most pronounced in the first quintile but remains significant in the remaining four quintiles. For example, 40 of the 48 pairwise comparisons for the last quintile are still in the direction of the system-neglect hypothesis ($p < .001$, two-tailed binomial test). Figure 4 also suggests that the learning that does take place occurs mostly in the highly precise conditions (*i.e.*, $d = 9$).

In the Online Supplemental Materials, Section C.4, we follow MW in further analyzing the pattern of reactions in Figure 3 by estimating a "Quasi-Bayesian" model to test for learning to detect change. This model is in the spirit of Edwards (1968) and others, who compared empirical responsiveness to Bayesian responsiveness in "bookbag and poker chip" tasks. This analysis formally rules out the possibility that the hypothesized pattern is an artifact of the specific sequences of signals. This analysis corroborates Figure 4 that most of the learning that occurs corresponds to less conservative responses to highly diagnostic signals and when there is a high base-rate of change.



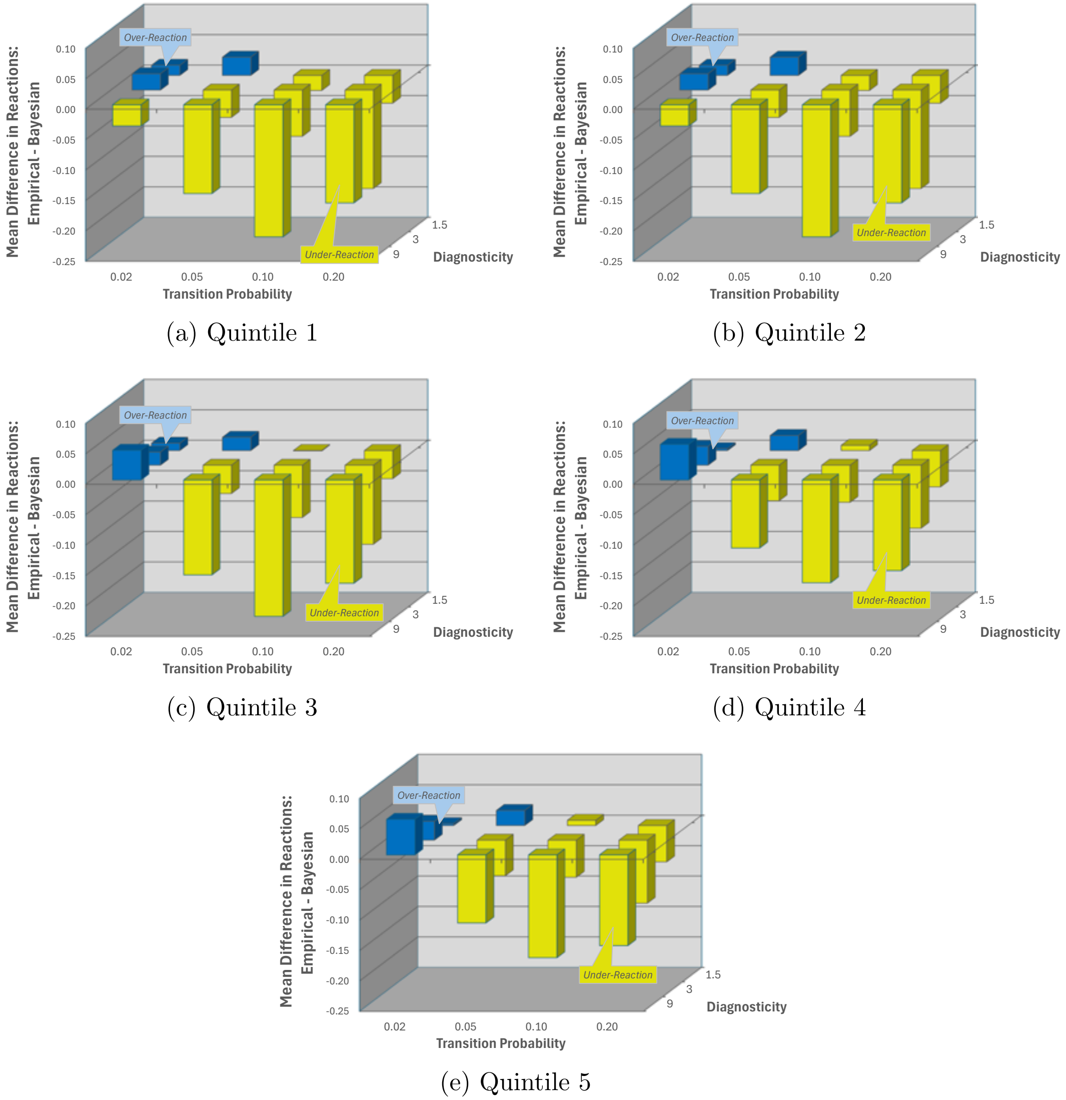
(a) Current Study

(b) Massey and Wu (2005a)

Fig 3. Over- and under-reaction to blue signals, by condition, as measured by the mean difference between the empirical reactions and Bayesian reactions, $p_i^e - \bar{p}_i^b$. The panel (a) shows this measure for the current study. Panel (b) shows the same measure for Massey & Wu (2005a)

5. The δ-ϵ model of adjustment

Applying Bayes' Rule to the change detection task is computationally complex and we do not expect participants to actually calculate Eqn. (1). In this section, we try to understand learning by viewing behavior through the lens of a computationally simpler and thus more psychologically plausible heuristic. First, we show that Bayes Rule for this task can be approximated with a simple linear adjustment heuristic, which we term the δ-ϵ model. The model adjusts the probability of a regime shift by δ with a signal indicative of change and by ϵ with a signal that is



(a) Quintile 1

(b) Quintile 2

(c) Quintile 3

(d) Quintile 4

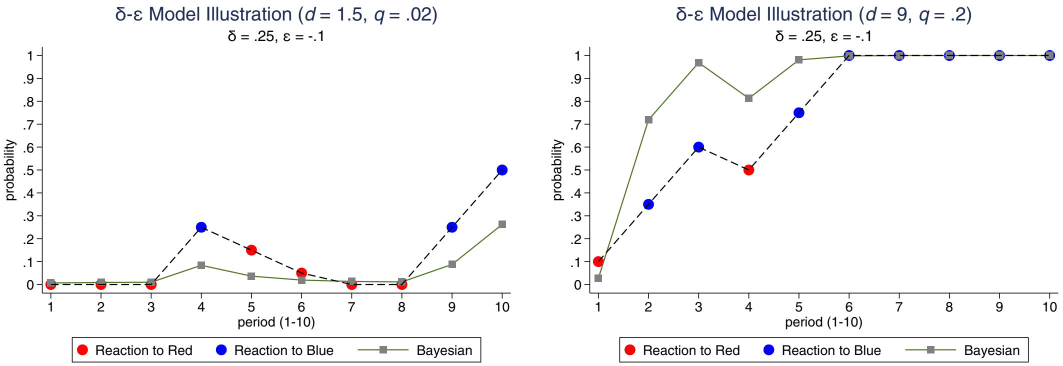
(e) Quintile 5

Fig 4. Over- and under-reaction, by condition and quintile (panels a-e), as measured by the mean difference between empirical reactions and Bayesian reactions, $p_i^e - \bar{p}_i^b$

indicative of no change. We then demonstrate how the δ-ϵ model fits the empirical probabilities provided by our 240 participants. Finally, we show how the δ-ϵ model provides insight into which systems are less exacting for "non-optimal" behavior and which systems provide more consistent feedback.

5.1. The δ-ϵ model

Recall that a trial consists of 10 signals, b_i, $i = 1, \ldots, 10$, where $b_i = 1$ for a blue signal and $b_i = 0$ for a red signal. Thus, the history of signals through i is given by $H_i = (b_1, \ldots, b_i)$, with $B_i = 1$ indicating that signal i was drawn from the blue bin.



(a) δ-ϵ adjustment for $d = 1.5, q = .02$ (b) δ-ϵ adjustment for $d = 9, q = .20$

Fig 5. Example of δ-ϵ adjustment that produces system neglect: over-reaction for (a) $d = 1.5$ and $q = .02$ and under-reaction for (b) $d = 9$ and $q = .20$, with $\delta = .25$ and $\epsilon = -.10$ for both systems

Let p_i indicate the probability that the ith signal was drawn from the Blue Bin, *i.e.*, the probability that the regime has shifted by signal i, $\Pr(B_i|H_i)$. For now, p_i refers to both normative (or, as we will show, approximately Bayesian) posterior probabilities, as well as empirical probabilities.

The δ-ϵ model is a linear adjustment model. We start with the first signal ($i = 1$):

$$p_1 = \begin{cases} q + \delta, & \text{if } b_1 = 1, \\ q + \epsilon, & \text{if } b_1 = 0. \end{cases} \tag{3}$$

In the absence of a signal, the probability of a switch is taken to be q, the transition probability. The probability adjusts by δ with a signal consistent with change, and ϵ with a signal consistent with no change, where $\delta \geq 0$ and ϵ may be positive or negative.

Adjustments for subsequent signals reflect the same linear adjustment, with an additional requirement that probabilities be bound between 0 and 1. In this model, the prior probability, p_{i-1} is a sufficient statistic for the history, H_{i-1}. We also investigated an alternative model in which blue signals "accumulate" (so that a red signal after many blue signals will not involve an adjustment), but we found that this model fits worse.

$$p_i = \begin{cases} \min(1, p_{i-1} + \delta), & \text{if } b_i = 1, \\ \max(0, p_{i-1} + \epsilon), & \text{if } b_i = 0. \end{cases} \tag{4}$$

We illustrate the model with two sequences from different systems, a system that generally produces under-reaction ($d = 9, q = .2$) and a system that generally produces over-reaction ($d = 1.5, q = .02$). Figure 5 illustrates how the model can produce system neglect. Here, we take $\delta = .25$ and $\epsilon = -.10$ for both systems. In the noisy and stable system, reactions should be relatively muted, whereas in the informative and unstable system, reactions should be more pronounced. The same reactions in the two systems clearly produce the predicted pattern of system neglect, with probabilities too high when $d = 1.5$ and $q = .02$ and too low when $d = 9$ and $q = .20$.

5.2. Correspondence to Bayesian probabilities

We next show that the δ-ϵ model can mimic Bayesian posteriors across all trials within a condition. To do so, we fit δ_b and ϵ_b to the Bayesian posteriors using a numerical grid search, in which we vary $\delta_b \in [0, 1]$ and $\epsilon_b \in [-1, .2]$ in increments of .01. The best-fitting parameters, δ_b^* and ϵ_b^*, minimized

Table 2 Fit of δ-ϵ model to Bayesian probabilities. The best fits minimize the sum of the squared deviations between the δ-ϵ and Bayesian probabilities, using a grid search. The deviation is the root mean squared error (RMSE) between the δ-ϵ and Bayesian probabilities

d	q	δ_{b}^{*}	$\epsilon_{\mathrm{b}}^{*}$	deviation
1.5	0.02	0.05	-0.01	0.034
1.5	0.05	0.09	-0.01	0.049
1.5	0.10	0.12	0.00	0.062
1.5	0.20	0.16	0.02	0.067
3	0.02	0.16	-0.13	0.050
3	0.05	0.22	-0.15	0.058
3	0.10	0.28	-0.15	0.068
3	0.20	0.29	-0.06	0.089
9	0.02	0.29	-0.27	0.041
9	0.05	0.39	-0.31	0.037
9	0.10	0.46	-0.19	0.049
9	0.20	0.58	-0.21	0.084

the sum of squared deviations between the model probabilities and the Bayesian probabilities for the 20 trials in each condition. Estimates that included the practice trials are nearly identical. The best fitting parameters are found in Table 2. Note that as d increases, δ_b^* tends to get larger and ϵ_b^* tends to get smaller. Note also that the fits are reasonably good as indicated by the root mean squared error (RMSE) from Bayesian probabilities, with RMSEs ranging from 0.034 to 0.089 across the 12 conditions.

In Figure 5, we took $\delta = .25$ and $\epsilon = -.10$ for two different systems. Table 2 shows that $\delta_b^* = 0.05$ when $d = 1.5$ and $q = .02$, with $\delta_b^* = 0.58$ when $d = 9$ and $q = .20$. Thus, $\delta = .25$ produces under-reaction in one system and over-reaction in the other system.[6]

5.3. Correspondence to empirical probabilities

We next fit the δ-ϵ model to each of our 240 participants, using the grid search approach outlined in the previous section, denoting the fits to empirical probabilities, δ_e and ϵ_e. The model was fit individually to each participant's 200 judgments for the 20 non-practice trials. For the most part, the model fits participants fairly well, with an average RMSE of 0.157 (range 0.000 to 0.410).

Figure 6 depicts a scatter plot of the estimates, δ_e and ϵ_e, along with the δ_b and ϵ_b for each condition. Note that this plot is consistent with system neglect. For $d = 1.5$ and $q = .02$, the δ_b and ϵ_b are northwest of most of the δ_e and ϵ_e fits, whereas the opposite holds for $d = 9$ and $q = .20$. We plot a measure of over- or under-reaction in Figure 7, the mean deviation between δ_b and δ_e. The pattern looks almost identical to that in Figure 3, with 43 of the relevant 48 pairwise comparisons consistent with system neglect ($p < .001$, two-tailed binomial test).

We also divide the data into quintiles and fit the model for each participant/quintile, δ_{ei} and ϵ_{ei}, where $i = 1, \ldots 5$. Figure 8 shows the estimates for Quintile 1 and Quintile 5. (Figure A.6 in the Online Supplemental Materials also depicts the other three quintiles.) We plot over- and under-reaction in δ, $\delta_b - \delta_e$ for Quintile 1 and Quintile 5 in Figure 9. Visually, it appears that under-reaction is less

[6] Note that we see the same pattern for ϵ. However, the system neglect hypothesis emphasizes reactions to signals of change relative to signals of non-change.

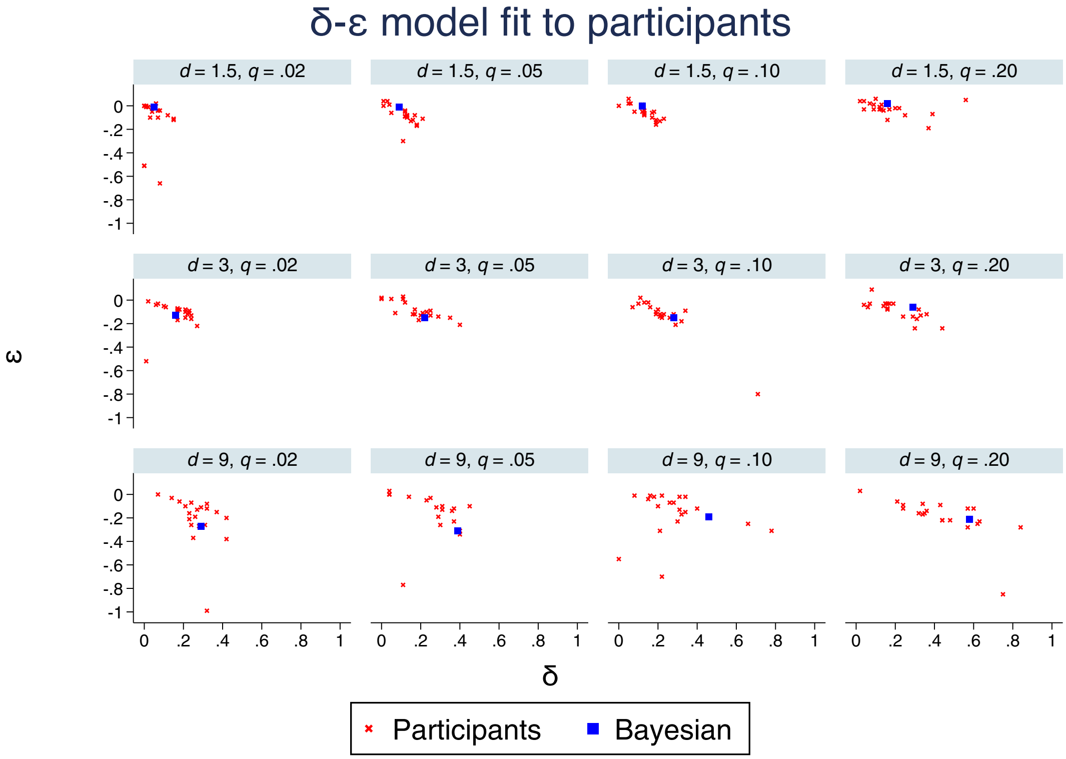


Fig 6. Estimates of δ-ϵ model fit to the 240 participants, by condition. The blue square shows the δ and ϵ for each condition that best fits Bayesian judgments

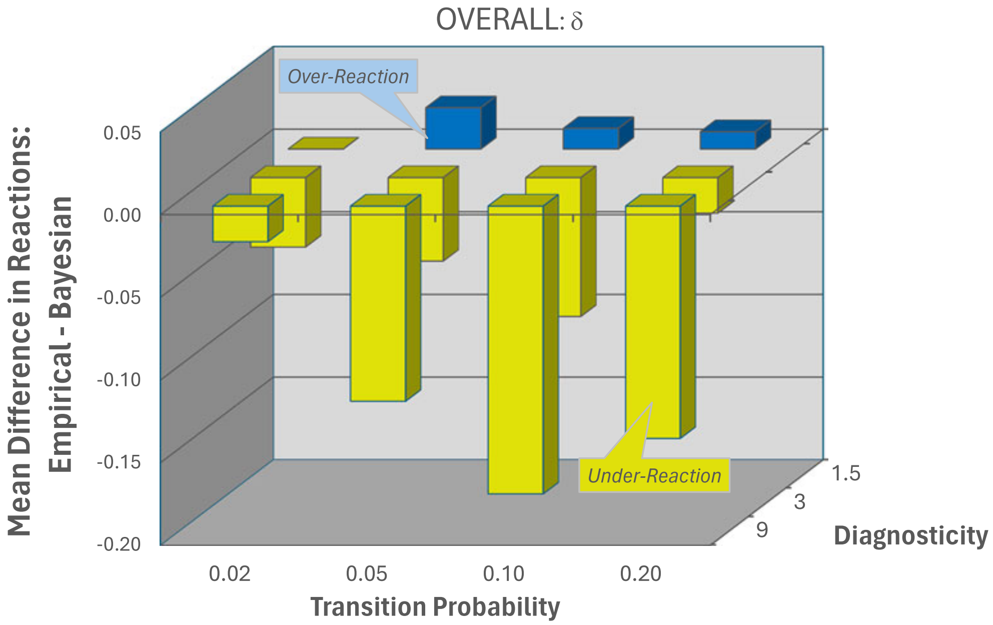
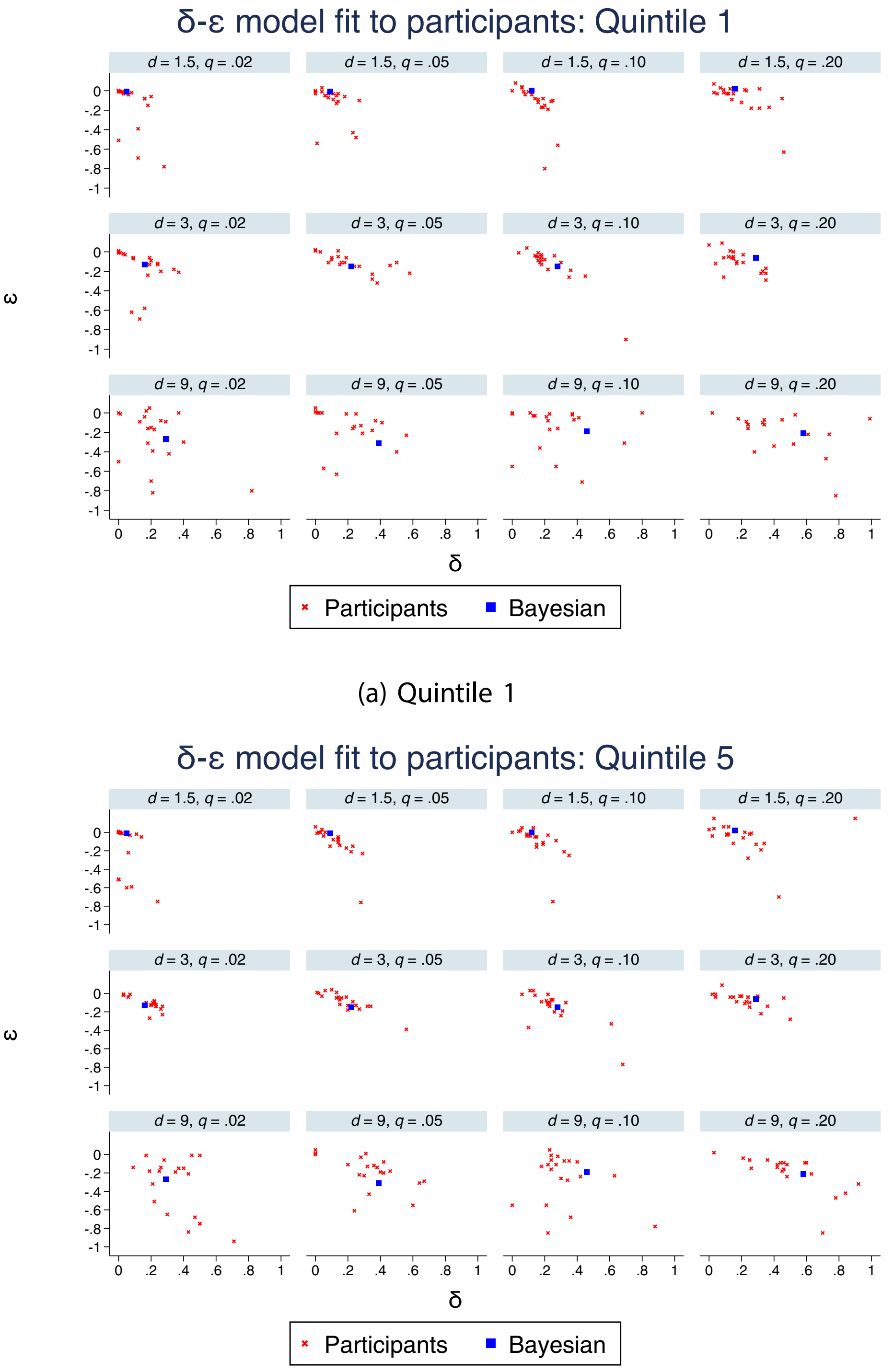


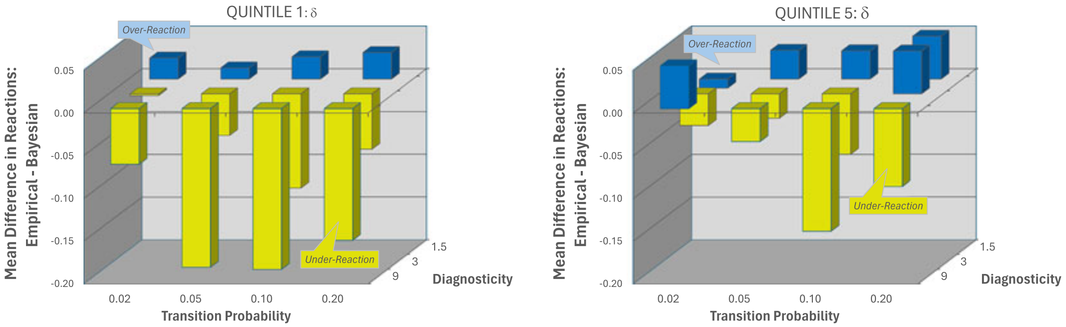
Fig 7. Over- and under-reaction to blue signals, by condition, as measured by the mean difference between the empirical and Bayesian δ, $\delta_e - \delta_b$



(a) Quintile 1

(b) Quintile 5

Fig 8. Estimates of δ-ϵ model fit to the 240 participants, by condition and for the (a) first and (b) fifth quintile. Model is fit to 40 judgments per participant and quintile. The blue square shows the δ and ϵ for each condition that best fits Bayesian judgments



(a) Quintile 1

(b) Quintile 5

Fig 9. Over- and under-reaction to blue signals, by condition and for the (a) first and (b) fifth quintile, as measured by the mean difference between δ_e and δ_b, where δ_e is fit for each participant-quintile

pronounced in Quintile 5, with much of the change happening in the $d = 9$ conditions, although over-reaction is slightly more pronounced. Indeed, 40 of the 48 relevant comparisons are in the predicted direction for Quintile 1, with 26 of the comparisons significant at the $p < .05$ level. In contrast, for Quintile 5, 38 of the 48 comparisons are consistent with system neglect, with only 19 significant at the $p < .05$ level.

5.4. Exactingness in the δ-ϵ framework

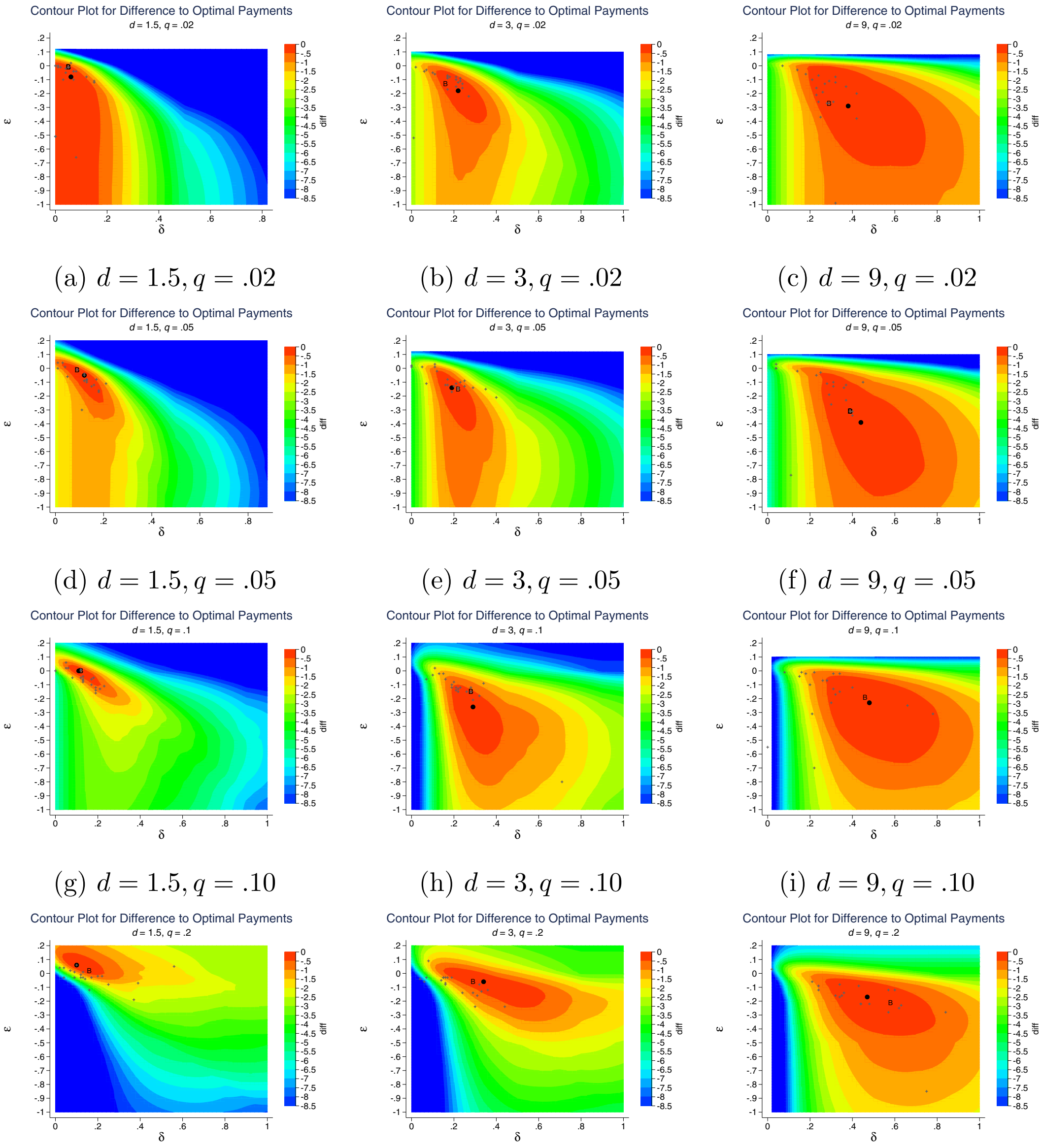
We have modeled reactions, both normative and empirical, in terms of linear adjustments to signals of change (δ) and signals of no change (ϵ). This δ-ϵ framework allows us to characterize each of our systems in terms of their conduciveness for learning.

We first examine how "exacting" earnings are to perturbations in δ and ϵ by looking at how earnings change when δ and ϵ parameters are non-optimal. For each condition, we calculate the overall earnings (*i.e.*, total earnings for the 20 trials of the study) that would accrue by varying $0 \leq \delta \leq 1$, and $-1 \leq \epsilon \leq 0.2$, in step sizes of .01. Let $\hat{\delta}$ and $\hat{\epsilon}$ denote the parameters that maximize total earnings for a condition. Let $E(\hat{\delta}, \hat{\epsilon})$, or $\hat{E}$ for simplicity, denote the total earnings for the earnings-maximizing set of parameters. For each condition, we compare earnings for combinations of parameter values, $E(\delta, \epsilon)$, to $\hat{E}$. We use a difference measure, $\hat{E} - E(\delta, \epsilon)$, as a measure of exactingness of earnings, but Figure A.9 in the Online Supplemental Materials depicts an alternative ratio measure, $E(\delta, \epsilon)/\hat{E}$, that provides nearly identical results. For example, for $d = 1.5$ and $q = .02$, earnings are maximized for $\hat{\delta} = 0.06$ and $\hat{\epsilon} = -0.08$, which produces a $E(\hat{\delta}, \hat{\epsilon})$ =15.03 $.Bycomparison$, E(.3, -.2) = 10.78, for a difference of \$4.25.

The contour plots in Figure 10 depict how exacting each system is to deviations from optimal levels of δ and ϵ. This analysis shows that systems vary considerably in exactingness. The systems with $d = 1.5$ (except for $q = .02$) are quite exacting in the sense that relatively small changes in reactions, δ or ϵ, can be costly in terms of performance. In comparison, the systems with $d = 9$, as well as $d = 1.5$, $q = .02$, are less exacting in the sense that relatively large changes in reactions lead to almost identical performance.

5.5. Consistency of best responses

In our experiment, participants received explicit feedback after each trial about when the regime actually shifted and therefore what series of probability judgments would have maximized earnings



(a) $d = 1.5, q = .02$ (b) $d = 3, q = .02$ (c) $d = 9, q = .02$

(d) $d = 1.5, q = .05$ (e) $d = 3, q = .05$ (f) $d = 9, q = .05$

(g) $d = 1.5, q = .10$ (h) $d = 3, q = .10$ (i) $d = 9, q = .10$

(j) $d = 1.5, q = .20$ (k) $d = 3, q = .20$ (l) $d = 9, q = .20$

Fig 10. Contour Plots for Differences between Maximum and Implied Earnings, $\widehat{E} - E(\delta, \epsilon)$, for different combinations of δ and ϵ in each condition (panels a-l). The circle in the middle of the bright orange section references the earning maximizing combination of δ and ϵ, $\hat{\delta}$ and $\hat{\epsilon}$. The "B" captures the δ and ϵ that best fits Bayesian posteriors. The gray $+$'s represent the δ and ϵ that best fit the 20 participants in that condition (see Section 5.3)

for that trial. Although this feedback is unambiguous, it was only provided at the end of each trial and therefore was not entirely generalizable. That is, knowing the actual regimes for a past trial allows for *a posteriori* optimal responses for that trial but may not lead to learning how to provide *a priori* optimal responses for future trials unless this feedback is relatively consistent. We therefore explore the consistency of trial-level feedback by looking at how a "best response" for one trial performs on a different trial.

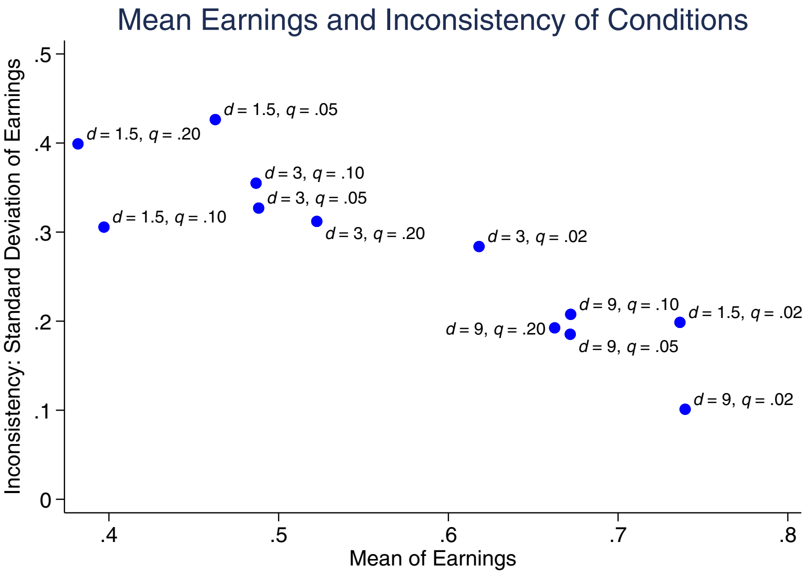


Fig 11. Mean and standard deviation of $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$ across the 12 conditions, where the standard deviation is our operationalization of (in)consistency

In our experiment, there are $s = 1, \ldots, 22$ series per condition.[7] For each series s, a participant saw a set of 10 signals, $H_s = (b_{1,s}, \ldots, b_{10,s})$. After the series was complete, participants were told if and when the regime changed. We denote that change as $\tau_s = 1, \ldots, 11$, where $\tau_s = 1$ indicates that the regime changes before the first signal was received and $\tau_s = 11$ indicates that the regime did not shift during that series.

For each of the series, $s = 1, \ldots, 22$, within a condition, we determine the earnings for using a δ-ϵ response strategy, $E_s(\delta, \epsilon)$. Let $\hat{\delta}_s$ and $\hat{\epsilon}_s$ denote the "best response" strategy, *i.e.*, the strategy that maximizes earnings given H_s and τ_s.[8] Our measure of consistency is $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$, where $s, s' = 1, \ldots, 22$, which measures earnings for using the best response for series s for series $s^{'}$.

To illustrate, consider a series in the $d = 9$, $q = .02$ condition, where $\tau_s = 7$ and $H_s = (1, 0, 0, 0, 0, 0, 1, 1, 0, 1)$. The best response to this series is $\hat{\delta}_s = 0.50$ and $\hat{\epsilon}_s = -0.27$, with $E_s(\hat{\delta}_s, \hat{\epsilon}_s) = \$0.75$. Using $\hat{\delta}_s$ and $\hat{\epsilon}_s$ on other series $s^{'}$ produces $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$ that ranges from \$0.48 to \$0.80, with a mean of \$0.76 and a standard deviation of \$0.07. Therefore, the best response for H_s performs consistently well across other series in that condition.

By contrast, $H_s = (0, 0, 1, 1, 1, 1, 1, 0, 1, 1)$ and $\tau_s = 2$ for a series in the $d = 3$,$q = .10$ condition. The best response to this series is $\hat{\delta}_s = 0.38$ and $\hat{\epsilon}_s = 0.20$, yielding $E_s(\hat{\delta}_s, \hat{\epsilon}_s) = \$0.22$. In this case, the best response for s often does poorly for $s' \neq s$, with $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$ ranging from \$-0.64 to \$0.77, with a mean of \$0.22 and a standard deviation of \$0.59.

Figure 11 plots the mean and standard deviation of $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$ across the 12 conditions. Note that these two measures are strongly negatively correlated ($\rho = -0.89$). The conditions in which best responses for one series s also work for other series $s^{'}$ (*i.e.*, high consistency) also generate high

[7] Note that we use the terminology series here instead of trial, because the series were randomized for each participant in order to correspond to a sequence of trials (though all participants in a condition saw the same two practice series in the same order). In addition to the 20 paid series, we include the two practice series because participants still received feedback for them. Analyses for only the 20 paid series are similar.

[8] Since every H_s and τ_s does not produce a unique optima, we put a small weight (.01) on the other 21 series in that condition so that we have a unique optima.

earnings across all series. On the contrary, the conditions in which best responses for one series may not work for other series tend to produce lower earnings.

We should note that our consistency analysis naturally folds in the exactingness analysis from the previous section. Feedback is effectively inconsistent if the best response strategy for one series is very different than that for another series *and* earnings are exacting (*i.e.*, punishing to deviations from best responses).

5.6. *Using consistency to explain learning*

The δ-ϵ framework allows us to characterize each of our systems in terms of their conduciveness for learning. We have modeled reactions, both Bayesian and empirical, in terms of linear adjustments to signals of change (δ) and signals of no change (ϵ).

In learning-friendly environments, the optimal adjustments in one situation are relatively similar to optimal adjustments in other situations. We hypothesize that this will affect learning. The more someone is reinforced for a particular behavior, the better off they are if that behavior is generally valuable. On the other hand, it will be difficult to learn when local optima tend to depart from global optima. Consider a golfer playing a course for the first time, trying to "figure out" the greens: how fast they are, how much break there is, etc. Some high-end courses maintain perfectly consistent conditions across all 18 greens, so what is learned on one green will apply to all greens. But that is rare. More often, there are some commonalities as well as some chance variations—closer mowing, dead patches, drainage challenges—that interfere with what a golfer can extrapolate across greens. We propose that, with change-point detection as with golf, the more consistent the rewards, the more conduciveness to learning.

To evaluate the impact of consistency on learning, we relate changes in participant performance to the (in)consistency of their experimental condition, which we operationalize as the standard deviation of the consistency function, $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$. We operationalize learning as the difference in relative earnings (Bayesian minus empirical) between a participant's first quintile (their first four paid trials) and their fifth quintile. In this analysis, we also consider each participant's "scope for learning," the difference between their first quintile earnings and the earnings that would be achieved by using the optimal δ and ϵ for their experimental condition. Scope for learning is important to control for as it is related to learning artifactually—*i.e.*, those who start poorly in a noisy process are more likely to improve simply due to regression to the mean. Figure 12 shows that scope for learning varies across experimental conditions, revealing another form of system neglect—while optimal δ-ϵ parameters are quite different across experimental conditions, participants' early behavior is relatively similar.

We regress the difference in relative earnings between Quintile 1 to Quintile 5 on consistency, while controlling for scope for learning at the participant level. As predicted, learning is strongly related to consistency ($\beta = 6.71, SE = 2.03, t = -3.31, p = 0.001$), with more learning occurring when consistency is high (*i.e.*, low standard deviation of $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$) and poorer when consistency is low (*i.e.*, high standard deviation of $E_{s'}(\hat{\delta}_s, \hat{\epsilon}_s)$). Learning was also positive related to scope for learning ($\beta = 0.25, SE = 0.09, t = 2.69, p = 0.008$), with more learning when there was more scope for learning.

6. Discussion

Intuition suggests that the system-neglect pattern will attenuate with experience, with judgments becoming more Bayesian, but learning about probabilistic tasks is generally difficult (Brehmer, 1980).

We examined learning across 12 conditions that varied in signal diagnosticity and system stability. Learning was more prevalent in highly diagnostic conditions. We also found that system neglect was partially attenuated by 20 trials of experience, especially for moderately and highly precise conditions, as well as for highly unstable conditions.

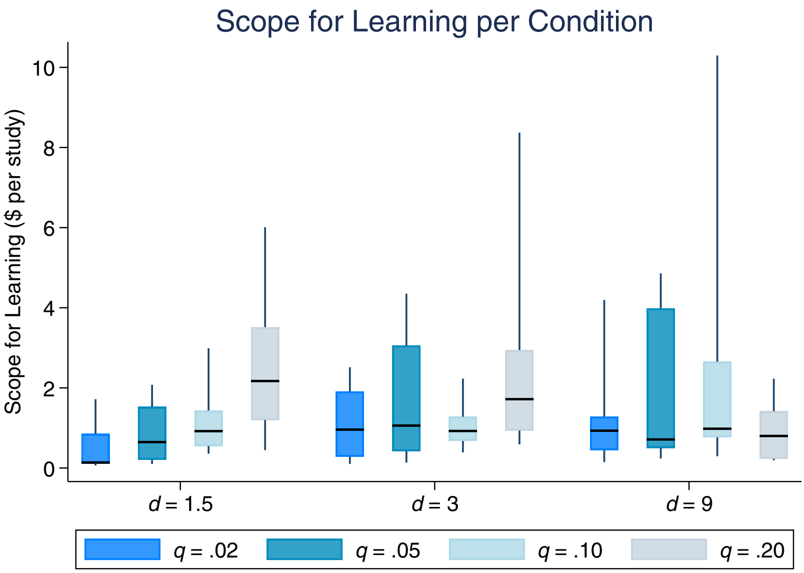


Fig 12. Box plot for scope for learning as measured by first quintile earnings relative to earnings that would be achieved by using the optimal δ and ϵ for their experimental condition. Scope for learning is rescaled to facilitate comparison with overall earnings for 20 trials. The box represents the interquartile range, with the whiskers representing the range from 10 to 90 percentile

To better understand this variation in learning across conditions, we examined two characteristics of the learning environments, corresponding to the exactingness of deviations from optimal adjustments (in δ and ϵ terms) and the consistency of feedback. We found more learning in conditions with environments that provided consistent feedback (for our tasks, such environments are also more exacting). In other words, it is hard to learn when the best response for one series might perform poorly on the next series.

We should caution that our analysis is limited to only 12 conditions, with three levels of diagnosticity and four levels of transition probability, and only 20 sequences per condition.[9] However, we see the theoretical analysis used in Figure 11 as generative in that it can be readily used to investigate other incentive schemes and conditions.[10] Although we used a quadratic scoring system in our experimental setup, our consistency analysis can be extended to make predictions about different payment structures. In Section C.9.4 in the Online Supplemental Materials, we show that the basic pattern in Figure 11 holds for linear payoff schemes, as well as in a prediction task in which participants provide a binary prediction about the current regime (as in MW Study 3). However, a payment scheme in which participants are penalized by deviations from the normative Bayesian standard more or less equalizes the conditions: now, all conditions provide highly consistent feedback. Although there is perhaps no real world analog to a “deviation from Bayesian” payoff scheme, an experiment with such a payoff structure would provide a sharp test of our theoretical analysis.

It is also straightforward to extend our analysis to other systems with different combinations of d and q. In Section C.9.2 in the Online Supplemental Materials, we simulate a large number of

[9] We did, however, find a remarkably similar pattern of results held for an earlier pilot study with only six conditions (the same three levels of diagnosticity but only $q = .05$ and .10). In that study, feedback inconsistency was the best predictor of learning.

[10] In Section C.9.1 of the Online Supplemental Materials, we repeat our consistency analysis for 500 simulated series for each condition. The same qualitative pattern obtains, though the negative correlation found in Figure 11 is somewhat attenuated.

systems with d ranging from 1.1 to 12.25 and q ranging from .005 to .52. Although mean and standard deviation of earnings are in general strongly negatively correlated as in Figure 11, this analysis identifies systems in which there is no correlation between mean and standard deviation. For the 12 systems used in our study, mean and standard deviation of $E_{s'}(\widehat{\delta}_s, \widehat{\epsilon}_s)$ are strongly negatively correlated. Indeed, the analysis we perform in Section 5.6, therefore, works as well with either measure as a covariate. We, however, have theorized that it is the consistency of feedback, as measured by standard deviation of $E_{s'}(\widehat{\delta}_s, \widehat{\epsilon}_s)$, that either facilitates or inhibits learning. A study that used systems with no correlation between mean and standard deviation would unconfound these two measures and help to pinpoint the mechanism for what is driving learning.

Finally, our analysis of consistency is based on the idea that the best δ and ϵ response for *one* series may or not be effective on *one* other series. Friedman's (1998) analysis of the Monty Hall Problem suggests that individuals do not aggregate unless prompted. He found that prompting participants to track cumulative earnings for both strategies (the commonly chosen but not optimal strategy, "Stick", as well as the optimal strategy, "Switch") significantly increased the number of participants who switched (see also).

Of course, in environments with low feedback consistency, aggregating over multiple series will yield more consistent feedback (see Online Supplemental Material, Section C.9.3). Counterintuitively, reducing the frequency of feedback might foster aggregation, akin to providing feedback every few trials as in Lurie & Swaminathan (2009).

6.1. Implications

While our study focused on detecting regime shifts in a relatively abstract context, our findings have implications for understanding how managers and policy-makers can improve their decisions in dynamic real-world contexts, such as pricing (Seifert et al., 2023) and stocking (Kremer et al., 2011). Depending on the decision environment, decision makers might under-react to early signs of system change, leading to delayed responses and potential losses. Conversely, they might over-react to minor fluctuations in performance metrics, resulting in unnecessary strategic shifts. Future research should examine whether decision makers improve their decision making with experience in such contexts, whether in laboratory simulations or with real-world data.

The influence of environmental factors on learning has direct implications for decision makers. Obviously, it would be helpful to increase the quality or quantity of feedback available to a decision maker. For example, instituting a waiting period before reacting to signals of change could help reduce feedback inconsistency and therefore rates of reacting to a false alarm. Unfortunately, firms often do not or cannot control the feedback available in their environment. However, they may be able to improve decision-makers' attention to feedback, by enhanced record-keeping or through activities explicitly aimed at learning from the past (Cyert & March, 1963, Friedman, 1998). Another approach is the use of policies to restrict decision-makers' freedom (Heath et al., 1998) with the goal of avoiding "noise chasing" (*i.e.*, over-reacting to inconsistent feedback). Both of these approaches—learning programs and policy-based decisions—are ways to improve institutional memory, an adaptive response to environments with inconsistent feedback. We should also note that although feedback in some conditions in our study was inconsistent, feedback was at least immediate, clear, and vivid, all qualities that facilitate learning (Nisbett & Ross, 1980, Hogarth, 2001, Maddox et al., 2003). Naturally occurring feedback, in contrast, is likely to be delayed, ambiguous, and obscure.

7. Conclusion

Our paper establishes the robustness of system neglect in people's change-point detection and demonstrates the relationship between characteristics of an environment and learning. In the end, we are somewhat sober about the ability of individuals to avoid systematic over- and under-reaction in non-stationary environments. However, we are also encouraged by the possibility of learning. Together these sentiments suggest that an important directions for future research is to understand how different decision environments impact the potential for learning to detect change.

Supplementary material. The supplementary material for this article can be found at https://doi.org/.10.1017/eec.2024.9.

Acknowledgements. It is extremely important since it's for a special issue dedicated to my deceased colleague. This paper is dedicated to Amnon Rapoport, whose papers (Rapoport and Burkheimer, 1973; Stein and Rapoport, 1978) and book (Rapoport et al., 1979) on this topic and broader work on dynamic decision making inspired the last two authors when they first started thinking about change point detection. The first author deeply misses Amnon, cherishing the memories of him as both a colleague and a friend. We also thank the guest editor, David Budescu, and the reviewers for their useful comments.

References

Atkeson, A. G., Kopecky, K., & Zha, T. (2021). Behavior and the transmission of COVID-19. *AEA Papers and Proceedings*, *111* (ay), 356–60. doi:10.1257/pandp.20211064.

Ball, L. (1995). Time-consistent policy and persistent changes in inflation. *Journal of Monetary Economics*, *36*(2), 329–350.

Barberis, N., Shleifer, A., & Vishny, R. W. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*(3), 307–343.

Barry, D. M., & Pitz, G. F. (1979). Detection of change in nonstationary, random sequences. *Organizational Behavior and Human Performance*, *24*(1), 111–125.

Benjamin, D. J. (2019). Chapter 2 - Errors in probabilistic reasoning and judgment biases. In B. Douglas Bernheim, Stefano DellaVigna, & David Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations 1*. (Vol. 2, pp. 69–186). North-Holland.

Blinder, A. S., & Morgan, J. (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit, and Banking*, *37*(5), 789–811.

Bolton, G. E., & Katok, E. (2008). Learning by Doing in the Newsvendor Problem: A Laboratory Investigation of the Role of Experience and Feedback. *Manufacturing and Service Operations Management*, *10*(3), 519–538.

Bowler, S., & Donovan, T. (1994). Information and opinion change on ballot propositions. *Political Behavior*, *16*(4), 411–435.

Brav, A., & Heaton, J. B. (2002). Competing theories of financial anomalies. *Review of Financial Studies*, *15*(2), 575–606.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, *45*(1-3), 223–241.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*(1), 49–67.

Budescu, D. V., & Hsiu-Ting, Y. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177.

Chinnis, J. O., & Peterson, C. R. (1968). Inference about a nonstationary process. *Journal of Experimental Psychology*, *77*(4), 620–625.

Chinnis, J. O., & Peterson, C. R. (1970). Nonstationary processes and conservative inference. *Journal of Experimental Psychology*, *84*(2), 248–251.

Coursey, D. L., Hovis, J. L., & Schulze, W. D. (1987). The disparity between willingness to accept and willingness to pay measures of value*. *Quarterly Journal of Economics*, *102*(3), 679–690.

Cyert, R. M. & March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ, Prentice-Hall.

Danz, D., Vesterlund, L., & Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, *112*(9), 2851–83.

Deming, W. E. (1975). On probability as a basis for action. *The American Statistician*, *29*(4), 146–152.

Donnell, M. L., & Du Charme, W. M. (1975). The effect of bayesian feedback on learning in an odds estimation task. *Organizational Behavior and Human Performance*, *14*(3), 305–313.

Edwards, W. (1968). Conservatism in human information processing. In Benjamin Kleinmuntz, (Ed.), *Formal Representation of Human Judgment* (pp. 17–52) New York, Wiley.

Erev, I. & Haruvy, E. (2015). Learning and the economics of small decisions. In John H. Kagel & Alvin E. Roth (Eds.), *The handbook of experimental economics* (Vol. 2, pp. 638–716). Princeton University Press.

Estes, W. K. (1984). Global and local-control of choice behavior by cyclically varying outcome probabilities. *Journal of Experimental Psychology: Learning Memory and Cognition*, *10*(2), 258–270.

Fader, P. S., & Lattin, J. M. (1993). Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science*, *12*(3), 304–317.

Friedman, D. (1998). Monty hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review*, *88*(4), 933–946.

Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Neglected risks: The psychology of financial crises. *American Economic Review*, *105*(5), 310–14.

Girshick, M. A., & Rubin, H. (1952). A bayes approach to a quality control model. *Annals of Mathematical Statistics*, *23*(1), 114–125.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435.

Grove, A. S. (1999). *Only the Paranoid Survive: How to Exploit the Crisis Points That Challenge Ever Company*. New York, Doubleday.

Guha, S., Seifert, M., & Ulu, C. (2024). The Role of Environmental Instability in Detecting and Responding to Signals of Impending Regime Shifts. In Federspiel, F.M., Montibeller, G., Seifert, M. (Eds.), *Behavioral Decision Analysis* (Vol. 350, pp. 105–120). Cham, Switzerland: Springer.

Hamilton, J. D. (2016). Chapter 3 - Macroeconomic regimes and regime shifts. In John B. Taylor, & Harald Uhlig (Eds.), *Handbook of macroeconomics* (Vol. 2, pp. 163–201). Elsevier.

Heath, C., Larrick, R. P. & Klayman, J. (1998). Cognitive repairs: How organizational practices can compensate for individual shortcomings. In Barry M Staw, & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 20, pp.1–37). Greenwich, CT, Jai Press.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists?. *Behavioral and Brain Sciences*, *24*(3), 383–403.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL, University of Chicago Press.

Hogarth, R. M., McKenzie, C. R. M., Gibbs, B. J., & Marquis, M. A. (1991). Learning from feedback: exactingness and incentives. *Journal Of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 734–752.

Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: system neglect and change detection. *Management Science*, *57*(10), 1827–1843 10.1287/mnsc.1110.1382.

List, J. A. (2003). Does market experience eliminate market anomalies?. *Quarterly Journal of Economics*, *118*(1), 41–71.

Lurie, N. H., & Swaminathan, J. M. (2009). Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes*, *108*(2), 315–329.

Maddox, W. T., Gregory Ashby, F., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Markus, B., Palia, D., Sastry, K. A., & Sims, C. A. (2021). Feedbacks: Financial markets and economic activity. *American Economic Review*, *111*(6), 1845–1879.

Martin, D. W., & Gettys, C. F. (1969). Feedback and response mode in performing a bayesian decision task. *Journal of Applied Psychology*, *53*(5), 413-418. doi:10.1037/h0028052.

Massey, C., & George, W. (2005a). Detecting regime shifts: The causes of under- and overreaction. *Management Science*, *51*(6), 932–947.

Massey, C. & George, W. (2005b). Electronic companion paper: 'Detecting regime shifts: The causes of under- and overreaction'. https://pubsonline.informs.org/doi/suppl/10.1287/mnsc.1050.0386

Nisbett, R. & Ross, L. (1980). *Human inference: Strategies and Shortcomings of Human judgment*. Englewood Cliffs, NJ, Prentice Hall.

Plott, C. R., & Zeiler, K. (2005). The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review*, *95*(3), 530–545.

Rapoport, A., & Burkheimer, G. J. (1973). Parameters of discrete time models of detection of change. *Management Science*, *19*(9), 973–984.

Rapoport, A. & Burkheimer, G. J., Stein, W. E. and Burkheimer, G. J. (1979). *Response Models for Detection of Change*. Dordrecht, Holland, D. Reidel.

Robinson, G. H. (1964). Continuous estimation of a time-varying probability. *Ergonomics*, *7*(1), 7–21.

Schweitzer, M. E., & Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, *46*(3), 404–420.

Seifert, M., Ulu, C., & Guha, S. (2023). Decision making under impending regime shifts. *Management Science*. 69(10), 6165–6180. doi:10.1287/mnsc.2022.4661.

Shewhart, W. A. (1939). *Statistical Method From the Viewpoint of Quality Control*. Washington, Graduate School, Department of Agriculture.

Sprecher, S. (1999). I love you more today than yesterday: Romantic partners' perceptions of changes in love and related affect over time. *Journal of Personality and Social Psychology*, *76*(1), 46–53.

Stein, W. E., & Rapoport, A. (1978). A discrete time model for detection of randomly presented stimuli. *Journal of Mathematical Psychology*, *17*(2), 110–137.

Steineck, G., Helgesen, F., Adolfsson, J., Dickman, P. W., Johansson, J. -E., Johan Norlen, B., Holmberg, L., & the Scandinavian Prostatic Cancer Group Study. (2002). Quality of life after radical prostatectomy or watchful waiting. *New England Journal of Medicine*, *347*(11), 790–796.

Theios, J., Brelsford, J. W., & Ryan, P. (1971). Detection of change in nonstationary binary sequences. *Perception and Psychophysics*, *9*(6), 489–492.

Wang, M. -C., George, W. & Shih-Wei, W. (2024) Detecting regime shifts: neurocomputational substrates for over- and underreactions to change. unpublished paper.

Yang, M., Han, C., Cui, Y., & Zhao, Y. (2021). COVID-19 and mobility in tourism cities: A statistical change-point detection approach. *Journal of Hospitality and Tourism Management*, *47*, 256–261.

Cite this article: Li, Y., Massey, C., & Wu, G. (2025). Learning to detect change: an experimental investigation. *Experimental Economics*, 1–26. https://doi.org/10.1017/eec.2024.9