Journal of Experimental Psychology: Learning, Memory, and Cognition

Searching for the Best Cause: Roles of Mechanism Beliefs, Autocorrelation, and Exploitation

Benjamin M. Rottman

Online First Publication, February 11, 2016. http://dx.doi.org/10.1037/xlm0000244

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000244

Searching for the Best Cause: Roles of Mechanism Beliefs, Autocorrelation, and Exploitation

Benjamin M. Rottman
University of Pittsburgh

When testing which of multiple causes (e.g., medicines) works best, the testing sequence has important implications for the validity of the final judgment. Trying each cause for a period of time before switching to the other is important if the causes have tolerance, sensitization, delay, or carryover (TSDC) effects. In contrast, if the outcome variable is autocorrelated and gradually fluctuates over time rather than being random across time, it can be useful to quickly alternate between the 2 causes, otherwise the causes could be confounded with a secular trend in the outcome. Five experiments tested whether individuals modify their causal testing strategies based on beliefs about TSDC effects and autocorrelation in the outcome. Participants adaptively tested each cause for longer periods of time before switching when testing causal interventions for which TSDC effects were plausible relative to cases when TSDC effects were not plausible. When the autocorrelation in the baseline trend was manipulated, participants exhibited only a small (if any) tendency toward increasing the amount of alternation; however, they adapted to the autocorrelation by focusing on changes in outcomes rather than raw outcome scores, both when making choices about which cause to test as well as when making the final judgment of which cause worked best. Understanding how people test causal relations in diverse environments is an important first step for being able to predict when individuals will successfully choose effective causes in real-world settings.

Keywords: causal reasoning, information search, dynamic environments, win−stay lose−switch

Supplemental materials: http://dx.doi.org/10.1037/xlm0000244.supp

The fundamental question of this article is how individuals test which of two causes produces the better outcome, such as choosing between two medications to reduce allergy symptoms, or two advertising campaigns to increase sales. Despite the obvious importance of being able to identify the better cause to achieve one's goals, it is a challenging task, complicated by factors such as how quickly the cause works and temporal trends in the effect. In the introductory sections, I frame the problem of searching for the best cause within the causal learning literature and as a type of sampling options associated with rewards more generally. Then I discuss challenges that arise due to temporal trends and causal mechanism such as delay and carryover effects. I present simulations to show how different testing strategies perform in different environments. Finally, I present the results of five experiments that examine whether individuals adapt their testing strategy to temporal trends and causal mechanism, and how they make a final decision of which cause works better.

Active Causal Strength Learning

This article focuses on a particular type of active causal strength learning: How to determine which of two causes works better by actively sampling the two causes repeatedly over time. Though there is a substantial literature on causal strength learning from observational (passive) experience (Cheng, 1997; Griffiths & Tenenbaum, 2005), and a growing set of literature on how learners actively make interventions for learning causal structure (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), there has been less focus on how individuals learn causal strength from active sampling. For example, a patient with chronic back pain tries to determine which of two medicines works best to alleviate his pain. Each day for 14 days the patient chooses one medicine to try, and at the end of the 14 days he decides which medication to use for the indefinite future. Active causal strength learning raises a number of previously unexplored challenges.

One reason why this task is so challenging is that, at any given time, it is only possible to know the outcome produced by the cause that was chosen. It is not possible to know how much pain the patient would have been in at time t had the patient taken Medicine 2 (or no medication) instead of Medicine 1 (Rubin, 1974, 1990; Splawa-Neyman, Dabrowska, & Speed, 1990). This counterfactual difference is the true difference in the effectiveness of the two medicines, but it cannot be directly measured.[1] Consequently, the learner must use temporal comparisons. For example,

Supported by the National Science Foundation (Grant 1430439) and the National Institutes of Health (Grant 1F32HL108711).

Correspondence concerning this article should be addressed to Benjamin M. Rottman, Department of Psychology, University of Pittsburgh, LRDC 726, 3939 O'Hara Street, Pittsburgh, PA 15260. E-mail: rottman@pitt.edu

[1] Splawa-Newman et al. (1990) used this counterfactual problem to motivate randomized controlled experiments, and Rubin (1974) used it to motivate propensity score matching for observational studies.

imagine the patient takes Medicine 1 on Day 1, and has a pain score of 56, and then takes Medicine 2 on Day 2, and has a pain score of 34. One option is to take the pain scores at face value and conclude that Medicine 2 works 22 points better than Medicine 1; however, a number of other interpretations are possible because Medicine 2 is tried after Medicine 1.

Before elaborating the complications that arise when searching for the best cause, I first briefly review research on how individuals actively sample noncausal options associated with rewards. Afterward, I resume discussion of the unique challenges of searching for the best cause.

Sampling Options Associated With Rewards and Similarity to Searching for the Best Cause

Psychologists have studied many cases of how people test options over time to determine which is associated with the best reward, often called *bandit problems*.[2] Bandit problems have some important similarities with and differences from searching for the best cause. I review these similarities and differences to help frame the unique aspects of searching for the best cause.

One dimension of bandit problems is whether they are static or dynamic. In dynamic problems, the rewards associated with each option change over time, and the goal is to exploit the currently best option, while periodically exploring other options to see if they have become dominant (Biele, Erev, & Ert, 2009; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Yi, Steyvers, & Lee, 2009). In contrast, in static bandit problems, the reward distributions associated with the options remain the same across time, so one can learn about an enduring difference between the options (e.g., Hills & Hertwig, 2010). Searching for the best cause is similar to a dynamic bandit problem in that the outcome (e.g., back pain) may change over time (e.g., after a period of low pain, there may a period of high pain due to an accident). However, in dynamic bandit problems, the best option can change over time; but when searching for the best cause, an intuitive assumption is that one cause is always better than the other, so it is possible to learn about an enduring difference of efficacy like in static bandit problems.

Another dimension of bandit problems is whether they just involve exploration (determining which option is best), or a combination of exploration and exploitation (selectively using the option believed to be best to try to increase one's rewards). The current studies focus on exploration-only tasks in which the goal is to try two causes for a relatively short period of time (14 samples) in order to decide which is best so that the chosen cause can be used for the indefinite future. The reason for this framing was to emphasize that the goal was to learn about an enduring difference in causal effectiveness.

Within all of these paradigms (static vs. dynamic, exploration-only vs. explore-exploit), one of the most important decisions a learner must make is the length of time to stick with one option before switching. Perseveration is the tendency to try one option repeatedly, whereas alternation is the tendency to switch between options. In one prototypical stable exploration-only task, the vast majority of participants switched less than 50% of the time, and about 50% of participants switched less than 20% of the time (Hills & Hertwig, 2010). In this study, there was a cognitive benefit to perseverating; it was associated with judgments more in line with expected utility. However, in dynamic environments, perseverating too much can be harmful. In one dynamic, task learners tended to stay with an option too long without realizing that another option had have become dominant (Yi et al., 2009). In another study, participants had a habit of performing the same action repeatedly even if it resulted in a bad outcome (Biele et al., 2009). The amount of perseveration versus alternation is also a critical determinant of success when searching for the best cause, which is discussed in the next two sections.

Tailoring a Preplanned Testing Strategy to Handle Autocorrelation

How should a learner decide to alternate or perseverate when searching for the best cause? When the outcome variable changes over time, such as if it comes and goes in waves, increases or decreases, or exhibits another secular trend, it is important to alternate. More generally, alternating is important when the effect is autocorrelated, when the pain at time t is correlated with the pain at time $t + 1$ (and less correlated with more distant time points), as opposed to being random from day to day. This principle is well known by single-subject research design methodologists, who advocate for "alternating" treatment designs (e.g., ABAB instead of AB, where A and B represent phases of different treatments), because alternating decreases the likelihood that the intervention pattern could be confounded with an underlying temporal trend (Barlow & Hayes, 1979).

Figure 1 makes this point with the example of trying two medicines over 14 days to determine which is best for reducing back pain. The solid line is the baseline pain the patient would experience without any medicine. The other lines represent the pain the patient would experience if the patient takes Medicines 1 or 2. Medicine 1 always works 5 points better than the Medicine 2, counterfactually; however, the patient only knows the amount of pain on a given day from the single medicine that was chosen, the circles and triangles.

In Figure 1a, comparing the average pain during Medicine 1 (Days 1–7) with the average pain during Medicine 2 (Days 8–14) implies that Medicine 1 works worse (higher scores). This incorrect inference is biased because the baseline generally decreases over time. In autocorrelated environments, perseverating can lead to large errors for or against either cause depending on whether the baseline function is increasing or decreasing and the order in which the causes are tried. In contrast, comparing the average pain scores after trying Medicine 1 versus Medicine 2 when alternating between the two (see Figure 1b) reveals the true difference in efficacy of five points, because both causes are tried at similar levels of the baseline function.

The bottom half of Figure 1 presents a baseline function that is random (autocorrelation low) from day to day. When autocorrelation is low it would be unlikely for the choice of medicine to be confounded with the baseline function. If Medicines 1 versus 2 happened to be taken on days with fairly high versus low baseline pain, this would be due to chance, so large errors in comparative

[2] Bandit problems are situated within a larger literature on active learning and information search (Gureckis & Markant, 2012; Markant & Gureckis, 2014; Meder & Nelson, 2012; Nelson, McKenzie, Cottrell, & Sejnowski, 2010; Nelson, 2005).

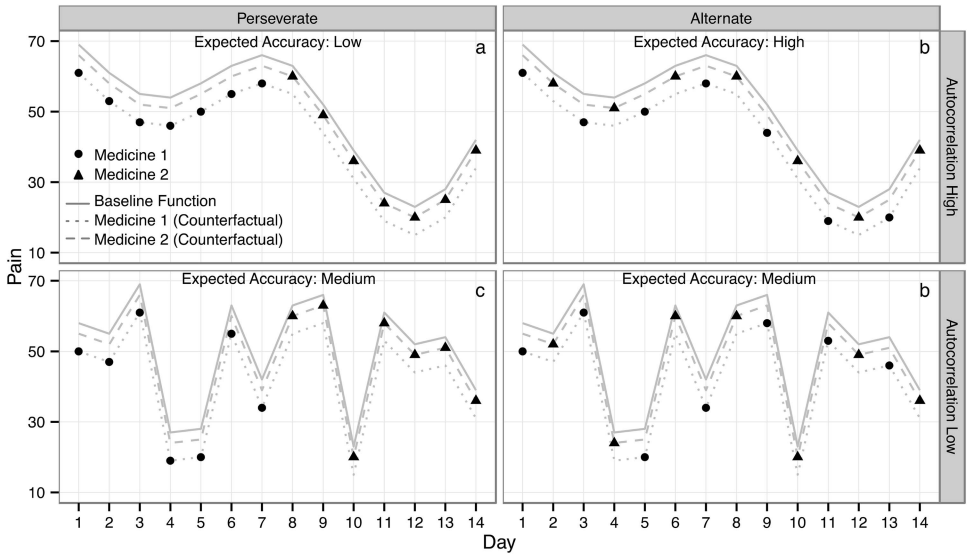


Figure 1. Example outcomes under two search strategies and two levels of autocorrelation.

efficacy would be unlikely. Furthermore, such a coincidence would be equally likely regardless of whether the learner alternates or perseverates. However, because of the noise in the baseline function, the expected accuracy of detecting the better cause is only moderate.

In sum, it is important to alternate when the baseline function is autocorrelated, but not when the baseline function is random over time. Later in the introductory sections I provide simulations to verify this intuition across various autocorrelated functions. The empirical question is whether individuals adapt their testing strategies based on the amount of autocorrelation in the baseline function.

Tailoring a Preplanned Testing Strategy to Handle Different Causal Mechanisms

There is yet another complication in assessing comparative efficacy from changes over time having to do with different mechanisms of how causes work. Tolerance is when a cause has decreasing effectiveness over time (e.g., caffeine, alcohol, and perhaps advertising campaigns work better when more novel). Sensitization is when a cause has increasing efficacy with repeated exposure (e.g., antidepressants, sensitivity to allergens, becoming sensitized to anxiety-provoking stimuli). Delay is when a cause takes time before working, and carryover is when a cause continues to work for a period of time after it is stopped. Tolerance, sensitization, delay, and carryover (hereon TSDC) are somewhat interrelated; Capturing exactly how a cause works with repeated dosage over time would require a mathematical model such as a pharmacodynamics model—see the next section for a simple simulation of delay and carryover effects. But the main point is that when testing causes that may have TSDC effects, it is important to try each cause for a period of time (perseverate) to allow each cause enough time to begin to work, to overcome any carryover effects, as well as to reveal tolerance or sensitization effects (Laska, Meisner, & Kushner, 1983).

People can quickly learn from experience whether a cause is exhibiting tolerance or sensitization effects (Rottman & Ahn, 2009), and there is a large body of research on how people use knowledge about delay for reasoning about causal relations (Buehner & May, 2002; Hagmayer & Waldmann, 2002; Lagnado & Sloman, 2006; Mendelson & Shultz, 1976). However, it is not known whether people use beliefs about TSDC effects when choosing a strategy to search for the best cause.

Simulations of Preplanned Testing Strategies for Identifying the Best Cause

The previous sections proposed, intuitively, why TSDC effects should lead a learner to perseverate, and why autocorrelation should lead a learner to alternate. These themes are addressed in the statistical literature on optimal experimental design of crossover trials (Bose & Dey, 2009; Jones & Kenward, 2003; Ratikowsky, Evans, & Alldredge, 1993; Senn, 1993); however, the statistical literature makes assumptions that do not fit well with the focus here on a single subject.[3] Consequently, I present simulations tailored to how well a learner can uncover the difference in the effectiveness of two causes. *R* code for these simulations is available from the author. Less interested readers can skip to the section Summary of Simulations of Preplanned Testing Strategies.

Each simulation (20,000 iterations) used a particular baseline function, and Cause 1 produced a 5-point increase relative to the baseline function, whereas Cause 2 did not make any change from baseline. The true difference in comparative effectiveness was coded as +5. Each simulation had 14 observations and compared three testing strategies. Perseveration was implemented as trying one cause for seven times and then the other seven times. Random testing was defined as trying both causes exactly seven times but

[3] These crossover designs assume that there are multiple arms or groups of subjects rather than one subject, and each group only receives a small number (e.g., 3 or 4) of treatment phases.

in a random order. Alternation was defined as switching back and forth between the two causes. Comparative effectiveness (E) was calculated as the difference between the average outcomes associated with Cause 1 minus the average outcomes associated with Cause 2.

Three statistics of E are reported. First, the mean of E across all iterations of a simulation (M_E) should ideally be +5. A score greater than 5 represents a bias that magnifies the apparent efficacy of Cause 1, a score less than 5 and more than 0 represents difficulty detecting the difference between causes, and a score less than 0 represents a bias that Cause 2 appears more effective than Cause 1. Second, the standard deviation of E (SD_E) is a measure of the precision of the estimate and is desired to be low. Third, $\%_{E>0}$ represents the percentage of iterations in which the simulation chose Cause 1 as the more effective cause (if $E > 0$) in a forced choice. It was very uncommon for E to be exactly zero. $\%_{E>0}$ is desired to be 100%. The larger that M_E is and the smaller that SD_E is, the more frequently the better cause will be identified. The following discussion focuses mainly on $\%_{E>0}$ for simplicity.

Autocorrelation in the Baseline Functions

Simulations 1–6 in Table 1 use six different baseline functions to verify the intuition that alternation is the best strategy in the face of positive autocorrelation. I chose parameters such that the functions have similar standard deviations; however, it is more meaningful to compare the three testing strategies within a function than to compare across functions. Asterisks denote the best strategy within a row for correctly identifying the more effective cause.

The unpredictable wavelike trend (UWT; see Equation 1) was developed as the baseline function for the experiments with the assumption that many variables change smoothly across time due to a gradual nonstationary process. Though the UWT is a sum of three sine waves, it only repeats every $156\pi \approx 490$ observations. Because participants saw 14 sequential observations chosen from a random starting position along the length (0, 156π), they saw many different trends (see Figure 2).

$$\text{UWT: } x_t = 50 + 20[\sin(t/3) + (2/3)\sin(t/2) + (1/2)\sin(10t/13)] \quad (1)$$

Simulation 1 in Table 1 shows that, for the UWT, it is easiest to identify the better cause when alternating, and the variance of the estimate of comparative effectiveness decreases considerably from perseveration to random testing, to alternation. Perseveration leads to very high error due to nonstationarity in the baseline function. This error can favor Cause 1 (e.g., see Figure 1a) or can favor Cause 2 (e.g., if Cause 2 was tried first in Figure 1a).

Simulations 2–5 examine other autocorrelated baseline functions, inspired by the autoregressive integrated moving average (ARIMA) model of time series analysis (Shumway & Stoffer, 2011). In the ARIMA model, three factors can lead to autocorrelation. The integrated component (I) of ARIMA corresponds to nonstationarity such as the UWT function and increasing or decreasing trends such as a random walk with drift (see Equation 2,

Table 1
Simulations of Preplanned Testing Strategies for Comparative Effectiveness

		Testing pattern								
		Perseveration			Random testing			Alternation		
Simulation	Details of simulation	M_E	SD_E	$\%_{E>0}$	M_E	SD_E	$\%_{E>0}$	M_E	SD_E	$\%_{E>0}$
Various baseline functions										
1	UWT	5	23	56	5	10	71	5	2	100*
2	Random walk with drift	5	37	50	5	12	67	5	5	75*
3	Autoregressive	5	16	63	5	8	76	5	4	91*
4	Moving average	5	13	65	5	10	71	5	4	91*
5	Negative autoregressive	5	5	85*	5	10	71	5	25	59
6	UTW randomized	5	10	71	5	10	71	5	10	71
Different kinds of delay and carryover; UWT function										
7	80%, 20%	5	23	56	4	10	67	3	2	96*
8	50%, 50%	5	23	55	2	10	60*	0	2	53
9	25%, 25%, 25%, 25%	3	23	54*	1	10	54*	0	2	52
10	20%, 60%, 20%	4	23	55*	1	10	52	−1	2	31
Delay and carryover; UWT randomized function (zero autocorrelation)										
11	80%, 20%	5	10	70*	4	10	67	3	10	63
12	50%, 50%	5	10	69*	2	10	60	0	10	50
Calculating comparative effectiveness with change scores (ΔNMH)										
13	UWT	1	9	53	5	5	86	10	1	100*
14	Random walk with drift	1	3	60	5	3	96	10	3	100*
15	Autoregressive	1	6	55	5	7	79	10	7	93*
16	Moving average	1	7	54	6	12	69	10	6	94*
17	UTW randomized	1	7	54	5	15	66	10	20	71*

Note. M_E is the mean of comparative efficacy; 5 is ideal. Scores greater than 5 (< 5) represent a systematically biased judgment in favor of the more (less) effective cause. SD_E is the standard deviation of comparative efficacy; 0 is ideal. $\%_{E>0}$ is the likelihood of choosing the better intervention; 100% is ideal. The delay and carryover numbers represent the percentage of the effectiveness of an intervention at Times t, $t + 1$, and $t + 2$. ΔNMH compares the average change score associated with Cause 1 versus Cause 2. ΔNMH = delta natural mean heuristic; UTW = unpredictable wavelike trend.
* Best testing strategy within a row according to $\%_{E>0}$.

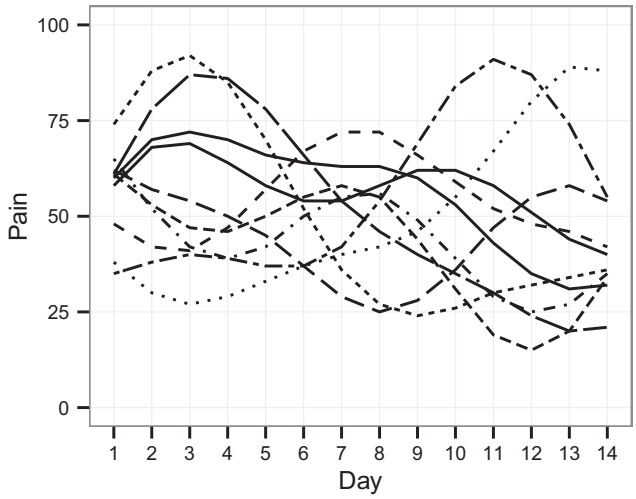


Figure 2. Ten example baseline trends from unpredictable wavelike trend function.

Simulation 2). Nonstationary baseline functions cause problems for perseveration because the testing pattern has a high likelihood of being confounded with the baseline function and, indeed, alternation was also the best strategy for the random walk with drift.

$$\text{Random walk with drift: } x_t = x_{t-1} + w_t;\ w_t \sim \text{Normal}(M = 5, SD = 5) \quad (2)$$

$$\text{Autoregressive: } x_t = .8x_{t-1} + w_t;\ w_t \sim \text{Normal}(M = 0, SD = 11) \quad (3)$$

$$\text{Moving average: } x_t = .8w_{t-1} + w_t;\ \ w_t \sim \text{Normal}(M = 0, SD = 15) \quad (4)$$

According to the ARIMA model, there are two other reasons for autocorrelation: the autoregressive (AR; see Equation 3) and moving average (MA; see Equation 4) components. The simulations show that the same pattern of results also holds for these functions. One obvious counterexample is negative autocorrelation (e.g., replacing the +.8 coefficient on x_{t-1} in Equation 3 with −.8). In a negative autoregressive function, the baseline trend alternates between high and low values. If the causes are tested by alternating, the causes would be somewhat confounded with the baseline function, which leads to high error in the estimate, so perseveration is better. Typically, time series modelers are more concerned about positive autocorrelation than negative, so hereon I use *autocorrelated* to mean *positively autocorrelated*.

The final baseline trend is the randomized UWT trend. This function was created by taking 14 sequential observations from the UWT trend, and then randomizing the order. The UTW randomized function has the same accuracy for all three testing patterns (as well as the UWT function [Simulation 1], using a random testing pattern) because the randomization in the trend and/or testing pattern guarantees that the testing pattern is independent of the baseline level in the long run.

The overall point of Simulations 1–6 is that, for a wide variety of positively autocorrelated functions (but not functions with zero or negative autocorrelation), alternation is better than random testing or perseveration.

Delay and Carryover Effects

Simulations 7–12 investigate how delay and carryover effects influence the estimation of comparative efficacy in the context of the UWT function.[4] The ARIMA framework uses transfer functions to describe the influence of a cause over time, such as how long it persists and when it has its maximum influence. Here I consider four specific transfer functions. In Simulation 7, a cause has 80% of its effect immediately (Lag 0) and 20% on the subsequent trial (Lag 1). Because Cause 1 produced a 5-point increase, for Simulation 7 it produced a 4-point increase at Lag 0 and the remaining 1-point increase at Lag 1.

Simulations 7–10 assume that the learner does not know the precise transfer function, and that comparative efficacy is calculated by comparing the average of the outcomes during Cause 1 minus the average outcomes during Cause 2—same as the previous simulations. With sufficient data, a learner may be able to infer the transfer function and use such knowledge when inferring comparative efficacy. However, the current article focuses on cases when the transfer functions are not known in advance of testing, and the purpose of these simulations is to show how different types of transfer functions could influence the choice of testing strategy.

Because the effectiveness is spread out over time, the accuracy of the comparative efficacy estimate diminishes (e.g., compare Simulation 1 with Simulations 7–9) because some proportion of the causal influence is attributed to the other cause. This reduction in accuracy occurs most dramatically for alternation because the accuracy for alternation was highest to begin with and because under alternation (compared with perseveration) a larger percentage of the causal influence gets attributed to the wrong cause. The usefulness of perseveration versus alternation is influenced by the percentage of the causal influence at Lag 0 versus Lag 1. When a large percentage of the influence occurs at Lag 1 (Simulation 10), alternation actually leads to systematically incorrect inferences, because the effect is attributed to the wrong cause. Simulations 11 and 12 repeat Simulations 7 and 8 using the UTW randomized baseline function to demonstrate that perseveration is best when there is *no* autocorrelation and when there are delay and carryover effects.

Summary of Simulations of Preplanned Testing Strategies

The simulations revealed the following points. First, with positive autocorrelation, alternation leads to the highest accuracy (Simulations 1–4). Second, when there is zero autocorrelation (Simulation 6), all three strategies are equivalent. Third, when there are delay and carryover effects and the baseline function is autocorrelated (Simulations 7–10), accuracy is fairly low and the best strategy depends on the amount of autocorrelation in the baseline function and the amount of delay and carryover. Fourth, when there are delay and carryover effects but autocorrelation is low (Simulations 11–12), perseveration is the best strategy.

These simulations and those below do not prove that one particular strategy is optimal. One could imagine a more sophisticated strategy that learns about the baseline function and the degree of TSDC and then adapts to those characteristics (though such a

[4] These simulations did not model tolerance and sensitization effects because there are so many possible patterns over time, and, in the presence of tolerance and sensitization effects, the concept of efficacy is vague. But intuitively, similar to delay and carryover, a cause must be tested enough times to assess whether it is having tolerance or sensitization effects.

model would be hard to devise in the first place and challenging to implement with only 14 observations). Rather, these simulations show that some strategies are more useful than others in different environments.

Alternative Ways to Infer Comparative Effectiveness

Another major question in this article is how learners estimate comparative efficacy. One option, sometimes called the natural mean heuristic (NMH; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig & Pleskac, 2008, 2010) and used in Simulations 1–12, compares the means of the outcomes during Cause 1 versus Cause 2.

In time series analysis, when the outcome variable is nonstationary, it is standard practice to take first- or second-order difference scores or "change" scores of the outcome variable to eliminate a linear or quadratic trend, before conducting further analyses. This is the integrated (I) component of ARIMA (Shumway & Stoffer, 2011). For example, consider a linearly decreasing baseline (20, 15, 10), and consider trying the causes in the order (2, 2, 1), when Cause 1 results in a score 5 points higher than Cause 2, resulting in (20, 15, 15). Calculating comparative efficacy from the raw scores incorrectly implies Cause 1 results in a lower outcome (15) than Cause 2 (17.5). In contrast, the difference scores calculated from (20, 15, 15) are (NA, −5, 0), and Cause 1 is correctly associated with a higher outcome (0) than Cause 2 (−5). (No difference score can be calculated for the first observation, which is represented by NA).

Simulations 13–17 are the same as Simulations 1–4 and 6, except that comparative effectiveness is calculated with change scores, which I call ΔNMH. Change scores improve the ability to identify the better cause for the two nonstationary functions, UWT and random walk with drift, for both random testing and alternation.[5] It has mixed effects for the autoregressive and moving average functions. Difference scores reduce accuracy in the UTW randomized function—there is no need to take difference scores with a random baseline function because they help account for nonstationarity. In the following studies I assess whether judgments about comparative effectiveness are predicted by NMH and or ΔNMH and whether learners adaptively use ΔNMH for autocorrelated environments.

Sequential Testing Strategies

To implement alternation or perseveration, a learner must plan the testing strategy in advance. Another possibility is that a learner sequentially decides which cause to test at each opportunity based on past experience with the causes. One motivation for sequentially testing causes is that a learner might think that testing the cause they currently believe to be more effective is actually the best way to assess comparative efficacy, a form of positive testing (Klayman & Ha, 1987). Another motivation is that a learner decides not only to learn about which cause is better (exploration), but also to attempt to produce the more desirable outcome at each opportunity by "exploiting" the cause that he or she currently thinks is better. The task in the experiments was described to participants as purely exploratory; however, in many real-world tasks, there is a need to balance exploration and exploitation. I discuss four sequential search strategies below.

One intuitive sequential strategy involves testing the cause that has the better average outcome from past experience. Sutton and Barto (1998) called this strategy "greedy search" (p. 28), and it is equivalent to NMH calculated sequentially after each observation to decide which cause to try next (Hau et al., 2008; Hertwig & Pleskac, 2008, 2010). I also implemented ΔNMH sequentially, which is the same as NMH, except it computes the mean outcomes on the change scores instead of the raw scores.

Another sequential strategy often used in explore−exploit tasks is win−stay lose−shift (WSLS), in which a learner repeats a choice if it previously resulted in a desirable outcome, and switches if it resulted in an undesirable outcome (e.g., Steyvers, Lee, & Wagenmakers, 2009; Worthy, Hawthorne, & Otto, 2013). In these studies, because the outcome is a number on the range 0–100 rather than binary, I used 50 as a cutoff for staying versus switching. I also implemented a version based on the change scores, called ΔWSLS (cf. Worthy & Maddox, 2014). In ΔWSLS, the agent stays with the same cause if there is a change in the favorable direction (e.g., a decrease in the outcome for pain), otherwise the agent switches.

Table 2 reports simulations of the four sequential strategies.[6] After the 14 testing choices are made, ΔNMH and ΔWSLS use ΔNMH for calculating comparative efficacy, and NMH and WSLS use NMH for calculating comparative efficacy. Table 2 reports three other metrics in addition to those in Table 1. "Alts" is the average number alternations. Because there were 14 trials, there is a maximum of 13 alternations. "Bal" (balance) is a measure of whether the causes were tried an equal number of times on a scale of 50% (balanced) to 100% (unbalanced). "Expt" (exploitation) is the percentage of times in which the more effective cause was tried during the 14 trials. NMH and ΔNMH result in fewer alternations, have higher imbalance, and exploit more than WSLS and ΔWSLS.

Across the four autocorrelated baseline functions, ΔWSLS had the highest probability of choosing the better cause, which can be explained by using change scores and by alternating the most. ΔNMH also has high accuracy for the two nonstationary functions. Though ΔWSLS has lower levels of exploitation than NMH and ΔNMH, it has higher levels than WSLS (in addition to higher accuracy). The raw score heuristics are better for UWT randomized for both accuracy and exploiting. Based on these simulations, one hypothesis is that learners will adapt to nonstationary auto-

[5] The change-score natural mean heuristic (ΔNMH) overestimates comparative effectiveness when alternating. For example, imagine alternating between causes (1, 2, 1, 2, . . .) on a flat baseline function (0, 0, 0, 0. . .) with the outcomes (5, 0, 5, 0 . . .) and difference scores (NA, −5, 5, −5 . . .); NMH = 5 but ΔNMH = 10. In addition, ΔNMH also underestimates comparative effectiveness when perseverating. For example, imagine trying the causes in the following order (1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2) with outcomes (5, 5, 5, 5, 5, 5, 5, 0, 0, 0, 0, 0, 0, 0) and difference scores (NA, 0, 0, 0, 0, 0, 0, −5, 0, 0, 0, 0, 0, 0); NMH = 5 but ΔNMH ≈ 1.

[6] Before these heuristics can make choices for subsequent trials, they require certain amounts of initial experience. Win−stay lose−shift (WSLS) only needs one trial to decide to stay or switch. ΔWSLS requires two trials so that there is one change score. The natural mean heuristic (NMH) requires one trial of each cause. ΔNMH requires one change score for each cause. To start the four heuristics off in the same way for comparability, the choice sequences were started with one of the four following options randomly selected (1,1,2), (2,2,1), (1,2,1), or (2,1,2), all of which satisfy the starting needs of all four heuristics. The models choose the causes for Trials 4−14.

Table 2
Simulations of Sequential Testing Strategies for Comparative Effectiveness

		NMH						WSLS					
Sim.	Details of simulation	M_E	SD_E	$\%_{E>0}$	Alts	Bal	Expt	M_E	SD_E	$\%_{E>0}$	Alts	Bal	Expt
18	UWT	11	17	81	4	82	72	10	16	73	6	70	59
19	Random walk with drift	12	23	73	5	79	69	7	24	67	7	71	53
20	Autoregressive	7	14	75	3	83	67**	7	12	74	6	68	57
21	Moving average	6	17	67	3	84	61**	7	13	71	6	65	56
22	UWT randomized	6	17	65	3	85	60**	5	11	70*	7	61	54
		ΔNMH						ΔWSLS					
18	UWT	8	7	93	3	85	81**	12	5	96*	7	69	67
19	Random walk with drift	5	5	89	2	86	79**	6	4	92*	7	66	61
20	Autoregressive	5	12	70	3	85	65	6	7	81*	7	61	57
21	Moving average	5	24	61	2	86	58	8	14	72*	7	61	56
22	UWT randomized	5	28	59	2	86	57	6	13	69	7	58	54

Note. M_E is the mean of comparative efficacy; 5 is ideal. Scores greater than 5 (< 5) represent a systematically biased judgment in favor of the more (less) effective cause. SD_E is the standard deviation of comparative efficacy; 0 is ideal. $\%_{E>0}$ is the likelihood of choosing the better intervention; 100% is ideal. Alts = average number of alternations with a maximum of 13; Bal = balance of whether one cause was tried more than the other (50% is perfectly balanced and 100% is totally unbalanced); Δ = change; Expt = percentage of trials in which the more effective cause was tried (exploited); NMH = natural mean heuristic; UTW = unpredictable wavelike trend; WSLS = win−stay lose−shift.
* Best testing strategy within all four heuristics according to $\%_{E>0}$. ** Best heuristic for exploiting.

correlated environments (e.g., UWT) by using change score heuristics but use raw score heuristics for randomized environments.

In this article, I test which of these four heuristics best explains individuals' sequential choices. I also examine whether learners adaptively switch strategies for autocorrelated versus randomized baseline functions.

Current Studies

Five experiments examine how people test which of two causes is more efficacious. Figure 3 summarizes the hypothesized learning and judgment processes. When choosing which cause to try at each testing opportunity, if a learner adopts a preplanned testing strategy, believing that TSDC is plausible should lead to more perseveration, whereas believing that the baseline function is autocorrelated should lead to more alternation. For learners who sequentially choose a cause to test based on prior experience, knowing that the baseline function is autocorrelated should lead to strategies based on changes rather than raw scores. When assessing comparative efficacy, ΔNMH should lead to higher accuracy than NMH for autocorrelated baseline functions but not for random baseline functions.

The goals of Experiments 1 and 2 were to examine causal testing behavior and assess the relation between testing behavior and accuracy of choosing the best cause in an autocorrelated environment. Experiments 3 and 4 tested whether subjects adaptively changed the search strategy and causal efficacy judgment calculation based on beliefs about autocorrelation and TSDC. Experiment 5 tested whether the findings would generalize to situations in which the learner chooses how much information to collect.

Experiment 1

Experiment 1 examined causal search behavior and how learners assessed which cause worked better in an autocorrelated environment. Unlike later experiments, participants were not given clear beliefs about TSDC effects and autocorrelation in the baseline function.

Method

Participants. A total of 152 participants (46% female) were recruited through MTurk.[7] Participants were paid $1, with the possibility of a bonus.

Stimuli and procedures. Participants read the following instructions:

> "Please imagine that you are a doctor treating patients for chronic back pain. There are two medicines that you can use. These medicines are meant to be taken once a day in the morning and they work all day. For 50% of patients Medicine 1 works better, and for 50% of patients Medicine 2 works better. Thus, you are going to try to figure out which medicine works the best for each individual patient. Every morning you decide whether the patient should take Medicine 1 or Medicine 2. Then you will see how much pain the patient is in during the afternoon. You have 14 days to test the medicines. At the end of the 14 days, you will judge which medicine works better and by how much. You will receive a bonus according to how close your estimate comes to the true difference in the effectiveness of the two medicines for the patient."

The bonus rate displayed was 20, 15, 10, 5, or 0 cents for a judgment within plus or minus 2, 4, 6, 8, or > 8 points of the true difference in comparative efficacy, respectively. Participants worked with eight scenarios, each of which represented a different patient, and were only told their bonuses at the end of the entire study.

Pain scores were presented numerically 0–100 (Figure 4a), *not* as a graph. One of the medicines, chosen randomly for each

[7] The intention was to recruit 100, but, due to a server error, 52 subjects were terminated early, and 52 more were recruited. All data were analyzed from all participants; omitting participants terminated early did not change any conclusions.

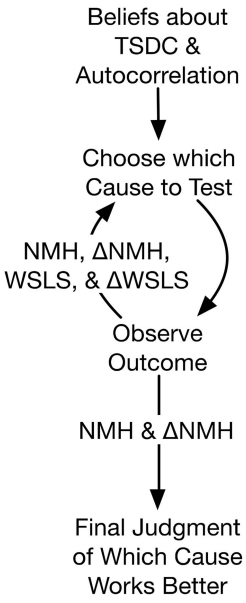


Figure 3. Hypothesized processes likely to influence causal testing and choosing the best cause. NMH = natural mean heuristic; WSLS = win−stay lose−shift.

patient, reduced pain by 5 points relative to the baseline; the other medicine did not change pain from the baseline. The 5-point difference was chosen to make the discrimination challenging but possible. The UWT (Equation 1) was used as the baseline trend, and for each of the eight scenarios the baseline trend had a different random starting position. Figure 2 shows 10 sample baseline trends to demonstrate the diversity that participants experienced. Participants never directly observed the baseline trend—they only observed the pain outcome after having chosen one of the medicines.

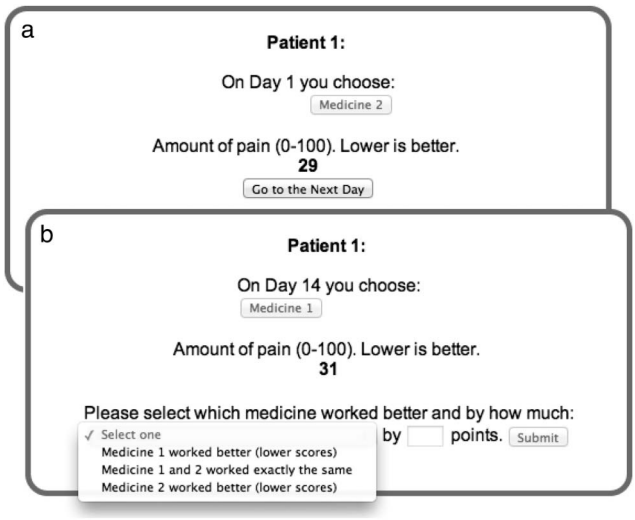


Figure 4. Screenshots of learning (a) and judgment (b) phases of Experiment 1.

After making the 14 choices and seeing 14 outcomes for a given patient, participants were asked to "select which medicine worked better and by how much" (see Figure 4b). Choosing that the medicines worked the same was coded as 0.

Results

Did people tend to alternate or perseverate? Figure 5 shows a histogram of the number of switches between medicines out of a maximum of 13, for each of the eight scenarios as well as a summary of all eight together. There are a number of important patterns in Figure 5. First, there appear to be three distinct strategies: alternating exactly once, alternating at every opportunity (exactly 13 times), or somewhere in the middle. Most instances of low ($n = 1$) and high ($n = 13$) alternations can be attributed to a relatively small percentage of participants (10 participants accounted for 68 of the 118 instances in which participants alternated exactly 13 times, and 26 participants accounted for 143 of the 212 instances in which participants alternated exactly once). Second, across the eight patients, there is a modest increase in the percent-

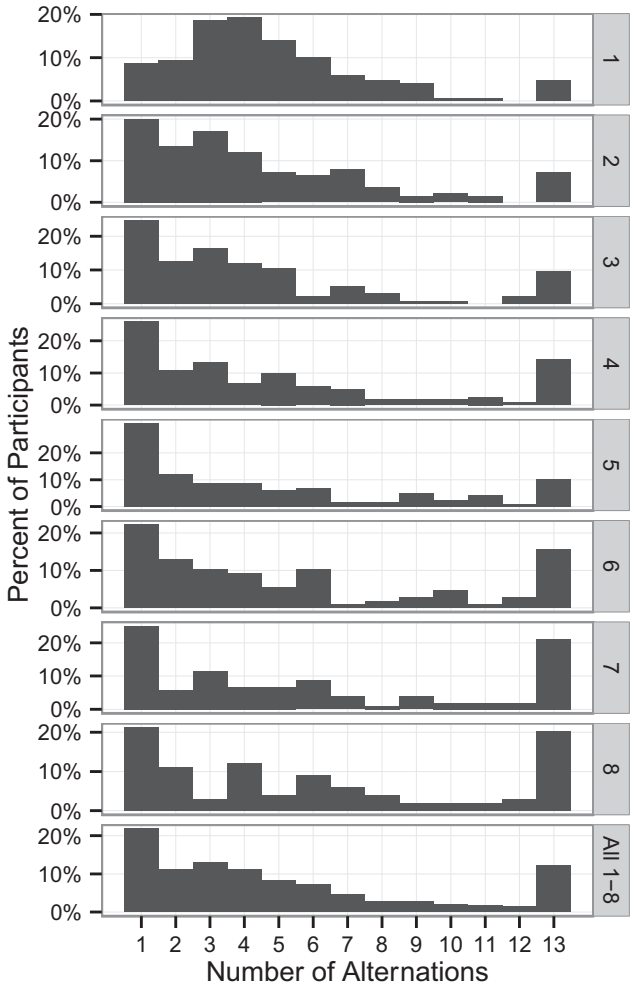


Figure 5. Histograms of the number of alternations by scenario in Experiment 1.

age of both extremes (exactly 1 and 13 alternations). Third, by far the majority of participants alternated fewer than seven times.

Lastly, there were 29 scenarios (out of 1,001 total) in which one medicine was tried for all 14 trials; six participants account for 22 of these cases. Because it is impossible to know which medicine works better if only one medicine was tried, these cases are not plotted in Figure 5 and are omitted from all analyses.

In sum, in the current task, most participants perseverated, but a minority frequently alternated. Because of the change in the distribution of alternations across patients, in Experiments 3–5 participants only worked with one patient case to study their initial search strategy.

Sequential testing strategies. I examined whether participants' choices of which medicine to test at each opportunity could be explained by the four sequential heuristics proposed in the introductory sections. I assumed that alternating exactly 13 times and trying one medicine seven times and then the other seven times were most likely preplanned strategies, so these scenarios (24%) were excluded from the analysis.

The four strategies—NMH, ΔNMH, WSLS, and ΔWSLS—were used to derive scores that predict the likelihood of staying with the same cause or switching to the other cause. For example, on any given trial, a score of 10 for NMH meant that the cause tried on that trial had a mean outcome 10 points higher than the mean outcome of the alternative cause. The heuristics predict that a score greater than 0 should lead to an alternation; in this analysis, continuous scores were used on the assumption that a higher score would more likely lead to an alternation than a lower score. Bivariate correlations between the four heuristics and the choice to stay versus switch are reported in Table 3; ΔNMH and ΔWSLS had the highest bivariate correlations. To make these correlations more interpretable, Table 3 also reports the likelihood of an alternation when each heuristic is positive versus negative. The probability of a switch when $\Delta WSLS > 0$ (48%) is over twice that of when $\Delta WSLS < 0$ (20%). The Appendix contains a figure of the relation between the continuous ΔWSLS score and the probability of alternation.

A multivariate logistic regression using standardized scores for all four heuristics was also run (see Table 3). The intercept, the base rate of alternation, was entered as a by-subjects random effect, and there were by-subject random slopes on the four heuristics, allowing for the possibility of variance in the use of the four strategies across participants. Three of the predictors were significant (with a multiple $r^2 = .08$). ΔNMH and ΔWSLS were also the strongest predictors in the multivariate analysis. This finding suggests that when deciding which cause to test, participants focused on changes in outcomes rather than raw outcomes, potentially to account for the nonstationarity in the baseline function.

Did people who alternated make better inferences? Accuracy of inference was assessed in two ways. The first accuracy measure was binary—whether the participant inferred the correct direction (e.g., Medicine 1 worked better than Medicine 2). For this analysis, instances in which participants inferred that Medicine 1 and Medicine 2 worked exactly the same were ignored. The second accuracy measure was log absolute error. For example, if Medicine 1 reduced pain by 5 points relative to Medicine 2 and a participant inferred that Medicine 1 increased pain by 20 points, the absolute error was 25 points. Log absolute error was used because perseveration can result in high error in favor of either the more or less effective cause, which appeared as high variance in the simulations.

Table 4 reports regressions that analyze the relations between alternation and error. The regressions had by-subject random effects for the intercept and by-subject random effects for the slope of number of alternations to account for repeated measures. A

Table 3
Comparisons of Sequential Testing Strategies for Choosing Which Cause to Test

	NMH	ΔNMH	WSLS	ΔWSLS
Experiment 1				
Bivariate	$r^2 = .02$ {.25 vs. .44}	$r^2 = .06$ {.24 vs. .46}	$r^2 = .006$ {.30 vs. .37}	$r^2 = .06$ {.20 vs. .48}
Multivariate	$b = 0.19$ (0.05),*** $\eta_p^2 = .001$	$b = .25$ (0.05),*** $\eta_p^2 = .01$	*ns*	$b = 0.72$ (0.07),*** $\eta_p^2 = .02$
Experiment 3				
Bivariate	$r^2 = .02$ {.40 vs. .54}	$r^2 = .009$ {.39 vs. .53}	$r^2 = .01$ {.43 vs. .52}	$r^2 = .01$ {.39 vs. .54}
Multivariate	$b = 0.29$ (0.11),*** $\eta_p^2 = .005$	*ns*	*ns*	$b = 0.20$ (0.08),* $\eta_p^2 = .002$
× Autocorrelation	*ns*	—	—	$b = 0.67$ (0.17),*** $\eta_p^2 = .008$
Experiment 4				
Bivariate	$r^2 = .005$ {.20 vs. .26}	$r^2 = .008$ {.18 vs. .28}	$r^2 = .01$ {.18 vs. .28}	$r^2 = .03$ {.16 vs. .32}
Multivariate	*ns*	*ns*	*ns*	$b = 0.50$ (0.11),*** $\eta_p^2 = .02$
× Autocorrelation	—	—	—	$b = 0.84$ (0.21),*** $\eta_p^2 = .008$
Experiment 5				
Bivariate	$r^2 = .002$ {.30 vs. .37}	$r^2 = .002$ {.30 vs. .36}	$r^2 = .003$ {.30 vs. .36}	$r^2 = .007$ {.26 vs. .36}
Multivariate	*ns*	*ns*	*ns*	$b = 0.22$ (0.07),** $\eta_p^2 = .002$
× Autocorrelation	—	—	—	$b = 0.40$ (0.14),*** $\eta_p^2 = .003$

Note. Standard errors of *b* coefficients are in parentheses. Values in curly braces {} are the probability of alternation when each heuristic is low versus high. NMH = natural mean heuristic; *ns* = not significant; dash = interaction not tested because the main effect was not significant; WSLS = win−stay lose−shift.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Relations Between Number of Alternations and Accuracy of Comparative Efficacy Judgments

Experiment	Autocorrelation high	Autocorrelation low	Autocorrelation Condition × Number of Alternations
Dependent variable: Choosing the better cause			
1	$b = 0.17$ (0.03),*** $r^2 = .09$	—	—
2	$b = 0.23$ (0.04),*** $r^2 = .13$	*ns*	$b = 0.24$ (0.04),*** $\eta_p^2 = .04$
3	$b = 0.26$ (0.06),*** $r^2 = .15$	*ns*	$b = 0.26$ (0.08),*** $\eta_p^2 = .03$
4	$b = 0.13$ (0.08), $p = .09$ $r^2 = .03$	*ns*	*ns*
5: Number of Alternations	$b = 0.13$ (0.06),* $r^2 = .06$	*ns*	$b = 0.15$ (0.07),* $\eta_p^2 = .02$
5: Percent Alternations	$b = 2.91$ (1.13),** $r^2 = .09$	$b = -1.17$ (0.66), $p = .08$, $r^2 = .03$	$b = 4.08$ (1.31),** $\eta_p^2 = .06$
Dependent variable: Log error of comparative efficacy			
1	$b = -0.09$ (0.01),*** $r^2 = .11$	—	—
2	$b = -0.05$ (0.01),*** $r^2 = .09$	*ns*	$b = -0.06$ (.01),*** $\eta_p^2 = .04$
3	$b = -0.11$ (0.01),*** $r^2 = .15$	*ns*	$b = -0.09$ (0.03),** $\eta_p^2 = .03$
4	$b = -0.06$ (0.04), $p = .08$, $r^2 = .02$	*ns*	$b = -0.10$ (0.05),* $\eta_p^2 = .02$
5: Number of Alternations	$b = -0.09$ (0.02),*** $r^2 = .24$	*ns*	$b = -0.11$ (0.02),*** $\eta_p^2 = .10$
5: Percent Alternations	$b = -1.48$ (0.28),*** $r^2 = .21$	$b = 0.68$ (0.26),** $r^2 = .07$	$b = -2.17$ (0.39),*** $\eta_p^2 = .14$

Note. Standard errors of b coefficients are in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

logistic regression found that more alternation was associated with a higher likelihood of choosing the better medicine, and a Gaussian regression found that more alternation was associated with less absolute logged error in the comparative efficacy judgment. Figures 6 and 7 visualize these analyses, with jitter on both axes to reduce overplotting. Judgments that the two causes were equivalent are plotted in Figure 6, but were not included in the logistic regression. Finally, given that the bonus was tied to accuracy, it is not surprising that an analogous regression found an increase of the bonus by .71 cents ($SE = .08$, $p < .001$, $r^2 = .11$) for every additional alternation, on average. Comparing a participant who alternates 13 times versus one time, this would result in a difference of 8.5 cents per scenario, and 68 cents for the entire experiment.

Heuristics for inferring comparative effectiveness. Table 5 reports bivariate correlations and multivariate analyses of NMH and ΔNMH for predicting the comparative efficacy judgment. The multivariate regressions were standardized and had a random effect on the intercept and slopes on both predictors. Both heuristics were significant in both analyses.

Discussion

Experiment 1 found that most participants alternated less than chance, and less alternation was associated with higher error in the ability to determine the best cause in an autocorrelated environment. The low rate of alternation could be viewed as a poor strategy for an autocorrelated environment; however, there is other evidence that subjects' testing strategies were somewhat adapted to the autocorrelated environment. For the sequential testing choices, ΔNMH and ΔWSLS explained subjects' choices the best of the four strategies. Likewise, the fact that ΔNMH explained variance in subjects' final comparative efficacy judgments, in addition to

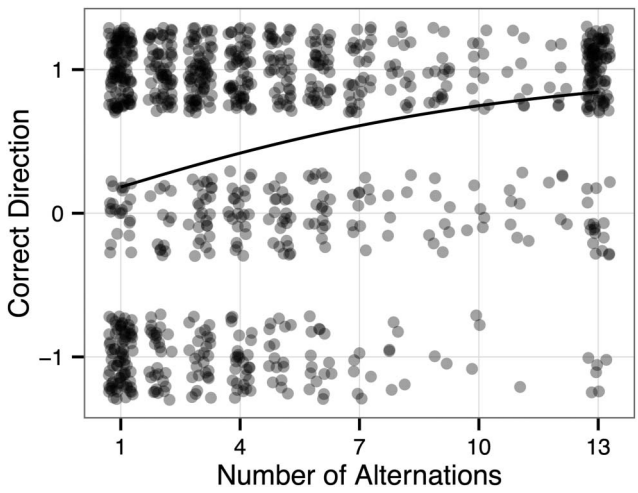


Figure 6. Judgments of which medicine worked better that were in the correct direction (1), incorrect direction (−1), and neutral (0), by the number of alternations in Experiment 1.

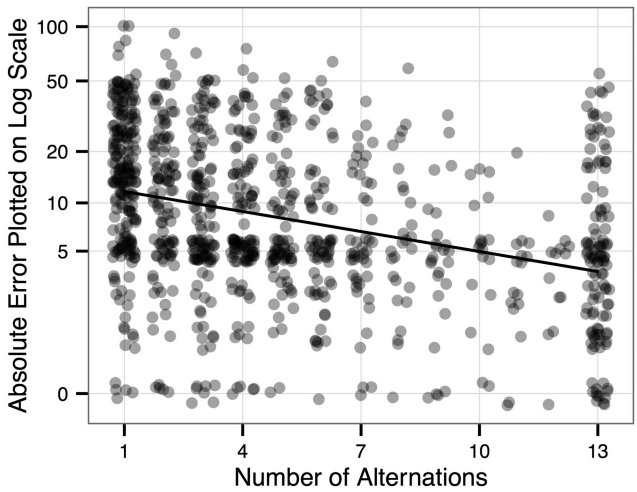


Figure 7. Errors of comparative efficacy judgments by number of alternations in Experiment 1.

Table 5
Results of Analyses for Inferring Comparative Efficacy

	NMH	ΔNMH
Experiment 1		
Bivariate	$r^2 = .21$	$r^2 = .28$
Multivariate	$b = 0.35$ (0.04),*** $\eta_p^2 = .11$	$b = 0.41$, (0.04),*** $\eta_p^2 = .19$
Experiment 2		
Bivariate	$r^2 = .30$	$r^2 = .19$
Multivariate	$b = 0.45$ (0.03),*** $\eta_p^2 = .21$	$b = 0.25$, (0.03),*** $\eta_p^2 = .08$
× Autocorrelation	$b = -0.17$ (0.08),* $\eta_p^2 = .005$	$b = 0.21$, (0.06),*** $\eta_p^2 = .02$
Experiment 3		
Bivariate	$r^2 = .12$	$r^2 = .08$
Multivariate	$b = 0.28$ (0.06),*** $\eta_p^2 = .07$	$b = 0.16$ (0.06),** $\eta_p^2 = .02$
× Autocorrelation	$b = -0.55$ (0.16),* $\eta_p^2 = .04$	$b = 0.74$ (0.13),** $\eta_p^2 = .10$
Experiment 4		
Bivariate	$r^2 = .29$	$r^2 = .24$
Multivariate	$b = 0.42$ (0.06),*** $\eta_p^2 = .21$	$b = 0.35$ (0.06),*** $\eta_p^2 = .16$
× Autocorrelation	$b = -0.31$ (0.12),** $\eta_p^2 = .04$	$b = 0.48$ (0.11),*** $\eta_p^2 = .09$
Experiment 5		
Bivariate	$r^2 = .27$	$r^2 = .09$
Multivariate	$b = 0.54$ (0.08),*** $\eta_p^2 = .19$	*ns*
× Autocorrelation	$b = -0.45$ (0.16),** $\eta_p^2 = .04$	$b = 1.14$ (0.23),*** $\eta_p^2 = .11$

Note. Standard errors of b coefficients are in parentheses. Δ = delta; NMH = natural mean heuristic.
* $p < .05$. ** $p < .01$. *** $p < .001$.

NMH, suggests that subjects were making use of change scores, which is beneficial in an autocorrelated environment.

Experiments 3–5 directly tested whether subjects adaptively changed their testing and causal inference strategies based on the amount of autocorrelation and TSDC beliefs. But before those studies, I conducted one more experiment to better understand the relation between alternation and accuracy.

Experiment 2

Experiment 2 manipulated whether subjects perseverated or alternated to better test the effect of alternation on accuracy. There are two motivations for this experiment. First, Experiment 1 found that alternation was associated with more accurate judgments of comparative efficacy. However, it is possible that people who tend to alternate happen to be better at this task, but that alternating itself does not cause inferences to be more accurate. This was tested in Experiment 2 by forcing participants to either alternate or perseverate.

Second, another explanation for the relation between accuracy and alternation is that alternating improves comparative efficacy judgments, but the improvement is due to a cognitive factor (e.g., alternation produces better memory or is easier to average), not that alternating produces cleaner data in an autocorrelated environment, as suggested by the simulations. To test this possibility, Experiment 2 compared the highly autocorrelated environment from Experiment 1 (UWT function) with an environment in which the baseline trend varied randomly from day to day (the UTW randomized function; hereon called the *autocorrelation low condition*).

The predictions of the simulations are that, in the autocorrelation high condition, alternation should lead to much higher accuracy than perseveration (100% vs. 56% accuracy in Simulation 1 in Table 1), but, in the autocorrelation low condition, there should be no difference (71% accuracy in Simulation 6). Deviations from this pattern, such as a benefit for alternation over perseveration or the reverse in the autocorrelation low condition, would suggest a cognitive benefit for alternation or perseveration.

In addition to these two primary motivations, Experiment 2 also tested whether subjects' final comparative efficacy judgments were based more on change scores (ΔNMH) for the autocorrelated function, and based more on raw scores (NMH) for the random function, which would be adaptive.

Method

Participants. One hundred participants (49% female) were recruited from MTurk. Twelve chose to stop participating before completing the study. I recruited up to 100 participants who fully completed the study, resulting in a total of 112 participants. All data were analyzed from all participants.

Design. The study design was 2 × 2 (Amount of Alternation [alternate vs. perseverate; between-subjects] × Autocorrelation [high vs. low; within subjects]). Participants were randomly assigned either to alternate (switch back and forth between the two medicines, resulting in 13 alternations across 14 days) or perseverate (try Medicine 1 for 7 days and then try Medicine 2 for 7 days).

The baseline trends were created in the following way. For each participant, four autocorrelation high trends were created by sampling 14 sequential data points from the UTW function. Then four parallel autocorrelation low scenarios were created by randomizing the order of the 14 data points within each of the four autocorrelation high scenarios. Thus, each participant saw four autocorrelation high and four autocorrelation low scenarios, but they have the same overall properties such as the mean and standard deviation.

Participants worked with all eight scenarios in blocks of autocorrelation high versus low scenarios. Half the participants received the high block first and half received the low block first.

Aside from these differences, Experiment 2 was the same as Experiment 1.

Results

Influence of alternation on accuracy of identifying the best cause. A logistic regression with random effects on the intercept and the slope of amount of alternation revealed that (a) participants who alternated were much more likely to infer the correct direction of comparative efficacy in the autocorrelated high condition, (b) there was no relation between alternation and accuracy in the autocorrelation low condition, and (c) the interaction between Autocorrelation and Alternation was significant. The same three effects were found for the log absolute error analysis using a parallel Gaussian regression. The regression results are presented in Table 4, and Figures 8 and 9; alternation was coded as 13 and perseveration as 1 so that the slopes of the regressions would be in the same units as Experiment 1. Means of the accuracy results across the four conditions are presented in Table 6; the means for the percent correct direction are close to the means predicted by the simulations (see the introduction to Experiment 2). Collectively, these results imply that the benefit of alternation is that it counteracts the effects of autocorrelation in the baseline function.

Heuristics for inferring comparative effectiveness. Just as in Experiment 1, both NMH and ΔNMH were used to predict the final comparative efficacy judgment. Bivariate correlations between the two heuristics and the comparative efficacy judgment revealed that both heuristics explained a considerable amount of variance (see Table 5). A multivariate analysis also revealed that both heuristics explained variance above and beyond the other (see Table 5); this model would not converge with random effects on the slopes of the heuristics, so only random effects on the intercept were used. In addition, interactions between the two heuristics and the autocorrelation condition revealed that the relation between NMH and comparative efficacy decreased from the autocorrelation high to low conditions, and the effect of ΔNMH increased (see Table 5). This finding fits with the idea that learners focus more on change over time in environments that are autocorrelated, but focus more on absolute scores in environments that are independent across time.

Discussion

The accuracy results imply that (a) there is a causal benefit of alternation on inferring comparative efficacy in autocorrelated environments, (b) this benefit is because autocorrelation reduces confounding with an autocorrelated baseline trend, and (c) in environments without autocorrelation, there is no inherent cognitive benefit for alternation or perseveration. This third point is interesting in that one might intuitively predict it is easier to calculate comparative efficacy in a perseveration condition, perhaps due to a cognitive cost of switching.

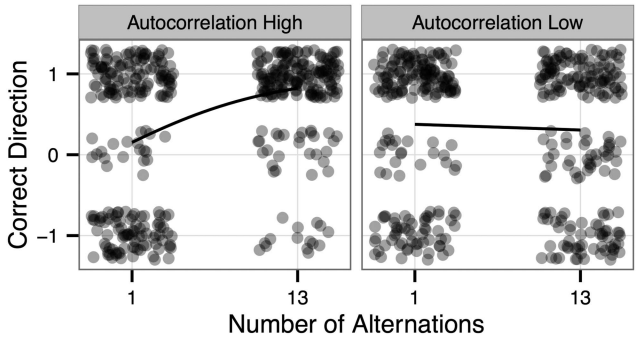


Figure 8. Judgments of which medicine worked better that were in the correct direction (1), incorrect direction (−1), and neutral (0), by the number of alternations in Experiment 2.

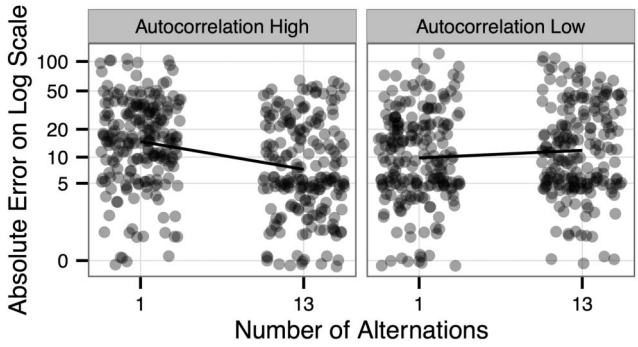


Figure 9. Errors of comparative efficacy judgments by number of alternations in Experiment 2.

Experiment 2 also found evidence of an adaptive shift in using the NMH more for estimating comparative efficacy in the autocorrelation low condition than in low high condition, and using ΔNMH more for the autocorrelation high than the autocorrelation low condition. The following experiments tested whether subjects used beliefs about TSDC and autocorrelation to guide their testing strategies.

Experiment 3

Experiment 1 found fairly low rates of alternation, despite the fact that alternation is beneficial in an autocorrelated environment. One explanation for this finding is that, going into the testing phase, subjects did not think that the baseline trend would be autocorrelated. Another explanation is that subjects thought that the medicines could have TSDC effects, in which case perseveration would be warranted.

The purpose of Experiment 3 was to test whether participants' background beliefs about autocorrelation and TSDC effects influenced their testing strategies. Believing that TSDC effects are plausible should lead a learner to perseverate more than believing that TSDC effects are implausible; perseveration would allow more time for a cause to become effective, for a carryover effect to dissipate, and more time to assess possible tolerance and sensitization effects. Believing the baseline function to have high as opposed to low autocorrelation should lead participants who preplan a testing strategy to

Table 6
Accuracy Results of Experiment 2

Autocorrelation	Alternate	Perseverate
	Percent correct direction	
High	91%	57%
Low	65%	68%
	Mean of absolute error	
High	13	22
Low	20	17

alternate more, and should lead participants who enact sequential strategies to focus more on change scores (ΔNMH and ΔWSLS) than on raw scores (NMH and WSLS).

In Experiment 3, autocorrelation and TSDC beliefs were manipulated by using cover stories for which participants would have preexisting beliefs that TSDC effects are plausible or not, and that the baseline function would be autocorrelated or not. The benefit of this approach is that subjects may be more likely to choose a testing strategy based on their own preexisting beliefs rather than on artificial manipulation. In contrast, if the cover story explicitly conveys information about TSDC and autocorrelation (Experiment 4), there is a higher risk of subjects choosing to ignore these aspects of the cover story if they find them artificial or incredible. However, one cost of manipulating the cover stories to tap into preexisting beliefs is that the stories are not perfectly matched pairs, which means that there is less control in Experiment 3. For this reason, a wide variety of different cover stories were used. In Experiment 4, beliefs about both autocorrelation and TSDC were manipulated within one cover story for higher internal validity.

Method

Participants. Three hundred participants (38% female) were recruited from MTurk, with about 20 participants per cover story. Participants were paid $1, with the possibility of the same bonus from the prior experiments.

Stimuli and design. Three different types of cover stories were created to activate beliefs about TSDC and autocorrelation in a fractional factorial design[8] (see Table 7). All stories involved testing two options to determine which option was associated with a better outcome. Participants had 14 opportunities to test the two options, and, at each opportunity, they had to decide to test Option 1 or Option 2.

Because the cover stories were not minimal pairs,[9] five cover stories were created for each of the three conditions so that each condition represented a range of scenarios. Using various cover stories also added variance in beliefs about TSDC and autocorrelation, which is useful for assessing whether different beliefs are associated with different search patterns.

The *base condition* had cover stories with high autocorrelation and low TSDC. In all these situations, the outcome variable (e.g., gas price, blueberry price, electricity generated by a solar panel, amount of perspiration, commute time) were plausibly highly autocorrelated. Furthermore, these scenarios did not involve causal interventions and thus TSDC effects were not plausible. (The one exception was that testing the deodorants in Story 4 did involve a causal intervention and it was at least conceivable that there could have been TSDC effects. This story was put in the base condition because TSDC effects seemed less plausible for deodorants than the interventions in the TSDC high condition.)

The *TSDC high* condition had stories with high autocorrelation and high TSDC. In all these scenarios, the outcome variable (back pain, allergies, vitamin D level in the blood, amount of thumb sucking, mood) were plausibly highly autocorrelated. Additionally, all involved causal interventions on one person over time, and thus TSDC effects were plausible.

The *autocorrelation low* condition had stories with low autocorrelation and low TSDC. In all these situations, the outcome variable should have been viewed as random from one observation to the next. Story 1 was about lottery tickets. In Stories 2–5, this randomness was accomplished by observing 14 independent people rather than following one person across 14 time points. Additionally, even though Stories 2–5 involved causal interventions, because each person was only observed once, it was not plausible that the interventions could have TSDC effects on the subsequent person who was observed.

Comparing the base case and TSDC high conditions tested whether beliefs about TSDC influenced the search strategy. Comparing the base case and autocorrelation low conditions tested whether beliefs about autocorrelation had an influence on search strategy.

Procedures. Participants were randomly assigned to one of the 15 cover stories (see Table 7). After reading the story, they answered two manipulation check questions about whether they believed the outcome variable would be autocorrelated or not, based on their prior knowledge. Participants were asked to imagine 14 observations (e.g., 14 days of back pain, 14 weeks of gas prices, 14 consecutive scratch-off lottery payoffs). Question 1, which used a 9-point scale, asked whether the outcome scores would be closely related to the prior observation (9), somewhat related (5), or unrelated to the prior outcome (1). The second question showed participants three graphs with low (A), medium (B), and high (C) autocorrelation, and participants were asked: "Do you think that the [outcome variable, e.g., pain] across the 14 [time periods] would look more like Graph A (1), B (5), or C (9)?"

Participants in the TSDC high condition were also asked to rate whether the causes would have TSDC effects. Participants were asked to imagine that they tried Cause 1 fourteen times (e.g., tried Medicine 1 for 14 days). Then they were asked the following four questions in order. With the exception of the deodorant story in the base condition, these four questions were not asked in the two TSDC low conditions because they did not make sense, and asking participants to make such a judgment could have encouraged bizarre beliefs about the scenario to accommodate the question. The questions were tailored to the specific cover story by inserting the correct name of the cause, be it a medicine, bribe with candy, yoga, and so forth.

> **Sensitization.** "How likely is it that [Cause 1] would initially have a small effect but would have a bigger and bigger effect over repeated use?" This question was rated on 9-point scale: 1 (*[Cause 1] probably*

[8] There is no fourth condition because it is difficult to conceive of situations in which participants, based on their prior knowledge, would strongly believe the baseline function to have low autocorrelation yet an intervention at one time could have some tolerance, sensitization, delay, or carryover (TSDC) effect at a later time. Typically, if TSDC effects are possible, such as in the one-person-over time scenarios, then it seems possible if not plausible that the background function could be autocorrelated. The goal for Experiment 3 was to rely only on participants' prior knowledge of the plausibility of autocorrelation and TSDC effects. In Experiment 4, beliefs about TSDC effects and autocorrelation were explicitly manipulated, allowing for this fourth condition.

[9] Comparing the tolerance, sensitization, delay, or carryover (TSDC) high condition versus base condition revealed why it is not possible to have one cover story fit all conditions. The TSDC high condition had stories that involve causal interventions, whereas those in base case did not have a causal intervention. The only way to have the same story in both conditions would have been to have a causal intervention but to stipulate through instructions that TSDC effects were not plausible—this is the approach of Experiment 4. But the goal for Experiment 3 was to allow participants to use their own beliefs about TSDC and autocorrelation.

Table 7
Synopses of Cover Stories in Experiment 3

Autocorrelation High/TSDC High
1. Testing two medications across 14 days to determine which was best at reducing back pain (same as Experiments 1 and 2).
2. Testing two medications across 14 days to determine which was best at reducing allergy symptoms.
3. Testing psychological reward versus punishment across 14 days to reduce thumb sucking in a child.
4. Testing two brands of vitamin supplement across 14 days to increase vitamin D in a patients' blood.
5. Testing yoga versus meditation across 14 days to improve a person's mood.
Base case: Autocorrelation High/TSDC Low
1. Choosing which gas station has lower prices after visiting one or the other for 14 weeks.
2. Determining which grocery store has lower blueberry prices after going to one or the other for 14 weeks.
3. Deciding which location on a roof to install a solar panel by testing how much electricity it generates in one or the other location over 14 days.
4. Choosing a new deodorant by trying one or the other for 14 days to see which results in the lowest amount of perspiration.
5. Choosing the faster route to work by trying one or the other for 14 days.
Autocorrelation Low/TSDC Low
1. Choosing between two instant (scratch-off) lottery games on 14 consecutive days to figure out which has the higher payoff.
2. Choosing which of two pain medicines works better by testing them on 14 different patients.
3. Choosing whether reward versus punishment works better to reduce thumb sucking in 14 children.
4. Having 14 consecutive restaurant customers taste and rate one of two brands of teas before deciding which to buy for future customers.
5. Choosing whether yoga versus meditation improves mood more in 14 separate patients.

Note. TSDC = tolerance, sensitization, delay, or carryover.

has the same effectiveness with repeated use), 5 (*Somewhat likely*), and 9 (*Very possible that effectiveness could increase with repeated use*).

Tolerance. "How likely is it that [Cause 1] would initially have a big effect but would have a smaller and smaller effect over repeated use? Note: you can answer 'very likely' for this question and the question above if you think that they are both plausible." This question was rated on 9-point scale: 1 (*[Cause 1] probably has the same effectiveness with repeated use*), 5 (*Somewhat likely*), and 9 (*Very possible that effectiveness could decrease with repeated use*). The tolerance and sensitization questions were written so that a participant could think that a cause might have a sensitization effect or a tolerance effect, but not have a specific belief about which one.

Delay. "How likely is it that [Cause 1] would take one or more days before starting to work—a delay?" This question was rated on 9-point scale: 1 (*Unlikely that there would be a delay*), 5 (*Somewhat likely*), and 9 (*Very likely that there would be a delay*).

Carryover. "You stop using [Cause 1] after the 14th day. How likely is it that [Cause 1] would continue to work for one or more days after stopping using it—a residual effect?" This question was rated on 9-point scale: 1 (*Unlikely that there would be a residual effect*), 5 (*Somewhat likely*), and 9 (*Very likely that there would be a residual effect*).

Next, participants were tasked with figuring out which of the two options produced a better outcome. Participants made 14 sequential choices between the two options. After they chose one option, they saw the outcome score (e.g., pain, price, minutes to work), and then made their next choice. The autocorrelation high and low baseline functions were the same as in Experiment 2, and participants made the same judgments of which option was better and by how much.

After the information search task, participants also rated the extent to which they were influenced by two motivations—exploiting and using a positive test strategy. The question about exploiting for Cover Story 1 in the TSDC high condition was: "When I thought that one [option, e.g., medicine] was working better than the other, I would continue to use that [option] **in order to reduce my level of** [outcome, e.g., pain] during the 14 days." The question about positive testing for this story was: "When I thought that one [option] was working better than the other, I would continue to use that [option] **in order to figure out whether it really works better or not to choose the best** [option] for the future." These questions were slightly reworded for each of the 15 cover stories to specify the options and the outcome, and to avoid stilted language (e.g., replacing "worked" with "had a higher. . ."). Exploiting and positive testing are discussed in the General Discussion. Other questions asked after the main task are documented in Supplemental Material 1.

Results

Nine participants were dropped from all analyses for never alternating. Figures 10 and 11 show histograms of the number of alternations collapsed by condition and by cover story, respectively, which will be discussed in the following sections.

Did autocorrelation beliefs influence alternation? Testing the influence of autocorrelation beliefs involved comparing the two TSDC low conditions (autocorrelation high vs. low). The two autocorrelation measures were averaged for conceptual simplicity (see Table 8 for means by condition); running all the analyses in this section separately for the two measures resulted in the same conclusions as when taking an average. Collapsing across cover stories, there was a significant difference in the autocorrelation judgments in the base condition (autocorrelation high/TSDC low) relative to the autocorrelation low/TSDC low condition, $t(202) = 7.01$ ($p < .001$, $d = 0.98$).[10]

In Figure 10, both conditions show a trimodal distribution, with groups of participants who alternated exactly once, exactly 13 times, and somewhere in the middle. Furthermore, there is no apparent shift toward more alternation in the autocorrelation high/TSDC low (base) condition compared with the autocorrelation low/TSDC low condition. However, examining each of the cover stories individually (see Figure 11) revealed that there are some different patterns in the amount of autocorrelation in the base condition; the deodorant cover story had low rates of alternation whereas the gasoline and blueberry cover stories had moderate to

[10] There was no difference between the two autocorrelation high conditions, $t(197) < 1$.

high rates. Because of the possibility that the base condition was not homogeneous, when testing whether autocorrelation beliefs had an influence on the amount of alternation, instead of just comparing the base condition with the autocorrelation low condition, I tested whether there was a correlation between autocorrelation beliefs and alternation within all 10 TSDC low conditions. For this analysis, the variability in the distributions in Figure 11 is not necessarily bad; it could be due to different beliefs about autocorrelation.

To assess whether participants who believed that the outcome had a higher degree of autocorrelation alternated more, I used multinomial regressions grouping the results as exactly one alternation, 2–12 alternations, or exactly 13 alternations, because of the trimodal distribution. The 2–12 strategy was set as the reference category, so the regression tested whether autocorrelation beliefs were associated with a change in the ratio of one versus 2–12 alternations, and a change in the ratio of 13 versus 2–12 alternations. The regression did not find any influence of participants' beliefs about autocorrelation on alternation ($ps > .73$). A linear regression also did not find an effect ($b = 0.11$, $p = .43$). In sum, even though participants had significantly different beliefs about autocorrelation across the scenarios, the amount of alternation did not track those beliefs.

Sequential testing strategies. It is possible that, instead of using more alternation in response to higher autocorrelation beliefs, participants modified their sequential testing strategies in response to autocorrelation, potentially by focusing more on change scores than on raw scores. As in Experiment 1, scenarios in which participants alternated every time or tried one cause seven times before trying the other cause seven times were omitted from this analysis, and only the two TSDC low conditions were used for comparability. Table 3 first reports bivariate correlations between each of the four sequential heuristics and the probability of alternating on each individual trial. All four heuristics correlated with the probability of alternation at roughly similar levels. As in Experiment 1, for each heuristic Table 3 reports the probability of an alternation when the heuristic is more than 0 and less than 0. For example, when ΔWSLS > 0, participants had a 54% chance of alternating, but when ΔWSLS < 0, the likelihood was 39%.

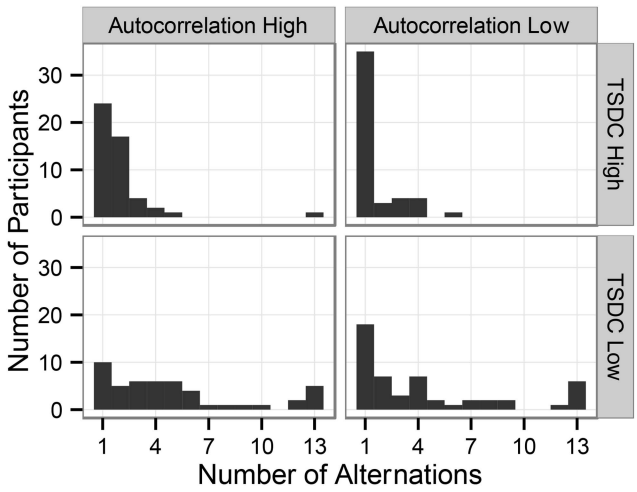


Figure 10. Histograms of the number of alternations by condition in Experiment 3. TSDC = tolerance, sensitization, delay, or carryover.

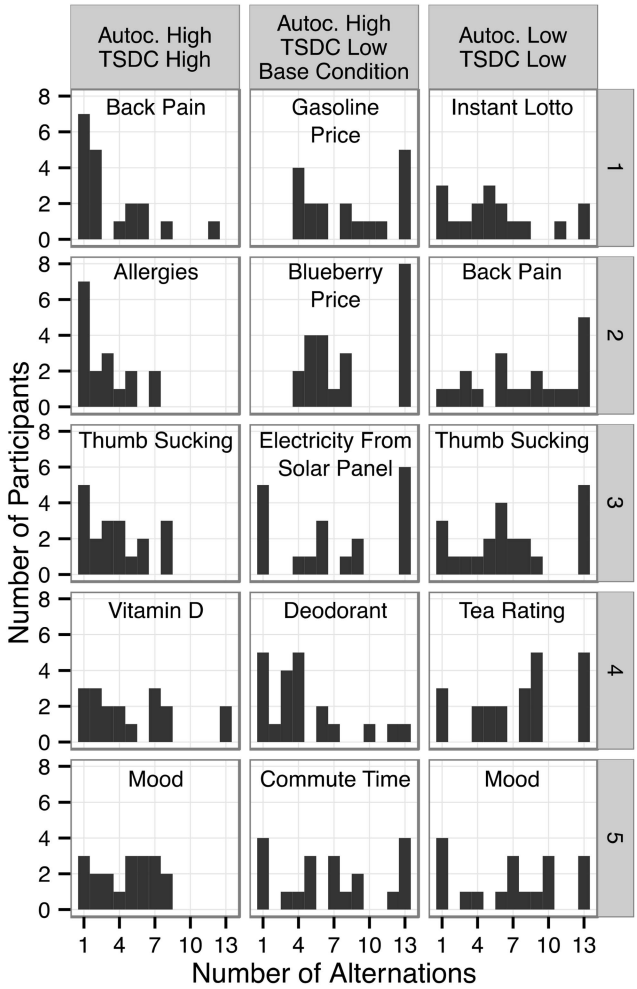


Figure 11. Histograms of the number of alternation by condition and cover story in Experiment 3. Autoc. = autocorrelation; TSDC = tolerance, sensitization, delay, or carryover.

Next, all four heuristics were entered into a multivariate logistic regression with a by-subject random intercept and by-subject random slopes for each of the heuristics to account for the fact that participants made 13 choices to alternate or not. Two predictors, NMH and ΔWSLS, were significant (see Table 3).

A follow-up regression was conducted to test whether NMH and ΔWSLS were moderated by the autocorrelation manipulation. Both the NMH, ΔWSLS, and two-way interactions with autocorrelation were entered into a logistic regression. (Due to convergence problems, only a random effect on the intercept was included.) There was no NMH $\times$ Autocorrelation interaction; however, the effect of ΔWSLS was higher for the autocorrelation high than the autocorrelation low condition (see Table 3). The interaction in these two regression curves for ΔWSLS when autocorrelation was high versus low is plotted in the Appendix. The variance in the ΔWSLS scores was much higher in the autocorrelation low than high condition; when autocorrelation was high, the changes in the outcome scores were fairly small; but when auto-

Table 8
Average Autocorrelation Ratings by Condition and Cover Story in Experiment 3

Cover story no.	Condition					
	Autocorrelation High/ TSDC High		Autocorrelation High/ TSDC Low (Base Condition)		Autocorrelation Low/ TSDC Low	
	M	*SD*	*M*	*SD*	*M*	*SD*
1	5.5	1.5	5.7	1.2	3.1	2.0
2	5.7	1.8	5.7	1.7	3.9	2.2
3	4.9	1.5	4.4	2.0	3.3	1.9
4	6.4	1.6	4.7	1.7	3.9	1.8
5	5.3	1.8	6.4	1.7	3.4	2.0
All	5.6	1.7	5.4	1.8	3.5	2.0

Note. Means and standard deviations are the average of the two autocorrelation judgments. The scales are from 1 (*low autocorrelation*) to 9 (*high autocorrelation*). TSDC = tolerance, sensitization, delay, or carryover.

correlation was low, the changes in the outcome scores from one observation to the next could be very high. What this graph shows is that, in the autocorrelation high condition, a decrease of 10 points was associated with alternating only about 35% of the time, whereas an increase of 10 points was associated with alternating about 55% of the time; a −10 versus +10 change is relatively dramatic, prompting a strong tendency to stay versus switch. In contrast, in the autocorrelation low condition, −10 and +10 changes are not uncommon, and these change scores were associated with alternations around 45% versus 50%, respectively (see Appendix).

In sum, there was evidence of use of both NMH and ΔWSLS for determining when to alternate and, furthermore, the use of ΔWSLS was moderated by autocorrelation in the data.

Did TSDC beliefs influence alternation? Answering the question of whether TSDC beliefs influenced alternation involves comparing the amount of alternation in the TSDC high condition with the base condition. In the TSDC high condition, almost all participants alternated less than chance (seven of 13 possibilities). In contrast, the base case (autocorrelation high/TSDC low) had a trimodal distribution (see Figure 10). Participants were moderately worried about the possibility of TSDC effects within the TSDC high condition. The average ratings were the following on a 9-point scale, with 5 being *somewhat likely*: tolerance ($M = 4.9$, $SD = 2.4$), sensitization ($M = 5.0$, $SD = 2.4$), delay ($M = 5.4$, $SD = 2.5$), and carryover ($M = 5.2$, $SD = 2.3$).[11]

Though there was a large difference in the pattern of alternations between the TSDC high versus base condition (see Figure 10), suggesting the expected pattern of less alternation with high beliefs in TSDC, there were two challenges in directly comparing these two conditions. The first, as already mentioned, is that there appeared to be heterogeneity in the base condition. The second problem is that participants also had higher ratings for exploiting and positive testing in the TSDC high condition than the base condition. Exploitation and positive testing (EPT) had a correlation of .70, and were averaged to create one composite measure. This composite measure was higher for the TSDC high condition ($M = 5.34$, $SD = 2.43$) than for the base condition ($M = 4.11$, $SD = 2.78$), $t(197) = 3.30$, $p = .001$. It is possible that the types of outcomes one might desire to change by taking some causal action are likely to be important in one's life (e.g., back pain, allergies, mood) and, consequently, susceptible to a desire to exploit.

Two strategies were used to test whether TSDC beliefs influenced alternation above and beyond EPT. First, a multinomial regression was run to predict number of alternations within these two conditions while controlling for EPT; the condition contrast and EPT were entered into the regression simultaneously. Higher scores on EPT where associated with a lower ratio of one versus 2–12 alternations ($b = -0.39$, $SE = 0.09$, $p < .001$),[12] and also a lower ratio of 13 alternations relative to 2–12 alternations ($b = -1.02$, $SE = 0.22$, $p < .001$).[13] The effect size for the total influence of EPT was $\eta_p^2 = 0.23$. Above and beyond EPT, the TSDC high condition compared with the base condition was associated with a higher ratio of one alternation relative to 2–12 ($b = 0.95$, $SE = 0.43$, $p = .03$), and was also associated with a lower ratio of 13 alternations relative to 2–12 ($b = -1.64$, $SE = 0.83$, $p = .05$). The effect size for the total influence of TSDC was $\eta_p^2 = .05$. In sum, subjects alternated less when they believed that TSDC effects were likely, controlling for the influence of EPT.

To test more specifically whether beliefs about TSDC influenced the amount of alternation, four regressions were run, one for tolerance, sensitization, delay, and carryover to predict the amount of alternation within the five TSDC high cover stories and also including the deodorant story from the base case, for a total of 113 participants. Poisson regressions were run because the TSDC high distributions were rightward skewed. None of the regressions were significant ($ps > .11$). In sum, these two analyses were at odds. Even though there was less alternation in the causal than the noncausal scenarios, there was no significant relation between subjects' beliefs about TSDC and alternation.

Did people who alternated make better inferences? There was (a) a negative relation between the number of alternations and log error for two autocorrelation high conditions, (b) no effect in

[11] In contrast, the deodorant story had lower average ratings, verifying that it belonged in the autocorrelation high/TSDC low condition: tolerance ($M = 4.6$, $SD = 2.8$), sensitization ($M = 3.4$, $SD = 2.6$), delay ($M = 2.4$, $SD = 1.9$), and carryover ($M = 3.0$, $SD = 2.4$).

[12] The reason for this shift is that exploitation and positive testing (EPT) often requires trying Option 1, then trying Option 2, and sometimes switching back to Option 1 if it is judged to be more beneficial than Option 1. In contrast, a common strategy that does not involve EPT is trying Option 1 for (roughly) 7 days and then Option 2 for 7 days.

[13] Alternating at every opportunity necessarily means not exploitation or positive testing.

the autocorrelation low condition, and (c) the Autocorrelation Condition × Alternation interaction was significant (see Table 4; see Supplemental Materials 2 for figures). These findings also hold when analyzing whether participants inferred the correct direction of which cause worked better (see Table 4 and Supplemental Materials 2). These findings replicated Experiment 2 in that alternation was only beneficial when the baseline trend was autocorrelated.

Heuristics for inferring comparative effectiveness. Bivariate correlations showed that both NMH and ΔNMH explained considerable amounts of variance is subjects' comparative efficacy judgments across all three conditions. A standardized multivariate analysis also revealed that each model predicted additional variance above the other. A follow-up regression was run to examine the two-way interactions between autocorrelation and the two heuristics. In the high as opposed to low autocorrelation condition, the effect of NMH was weaker and the effect of ΔNMH was stronger (see Table 5).

Discussion

Experiment 3 found no effects of prior beliefs about autocorrelation on the amount of alternation across a wide variety of cover stories. Furthermore, the fact that the amount of alternation in the autocorrelation high/TSDC low scenarios was only moderate was troubling because subjects could have performed better if they had alternated more, and, in this experiment, unlike Experiment 1, they knew ahead of time that autocorrelation was plausible and TSDC effects were implausible.

However, subjects adapted to the autocorrelation in other ways. For choosing which cause to test at each opportunity, participants were more sensitive to ΔWSLS in the autocorrelation high condition. Experiment 3 also replicated the finding from Experiment 2 that subjects shifted toward using ΔNMH more and NMH less for calculating the final comparative efficacy judgment when the baseline function was autocorrelated. These findings suggest that the adaption to autocorrelation occurred reactively, through direct experience with the data, rather than proactively.

The TSDC cover stories that involved causal interventions had much lower alternation than the other cover stories. On the other hand, within the causal intervention stories, there was no relation between TSDC beliefs and alternation. This raises the possibility that perseveration is a strategy that is used by default in causal contexts, like a script, but that a learner's prior knowledge about specific details of the causal mechanism is not necessarily used to guide the search.

Though a strength of Experiment 3 was that the cover stories drew out participants' own beliefs about TSDC and autocorrelation, the weakness was that the conditions were not perfectly matched pairs. Experiment 4 manipulated autocorrelation and TSDC beliefs directly using the same cover story, potentially clearing up some of the ambiguous results in Experiment 3.

Experiment 4

Method

Participants. Two hundred one participants (54% female) were recruited from MTurk. They were paid $1, with the possibility of the same bonus from the prior experiments.

Stimuli and design. Only the back pain cover story was used. The design was 2 × 2 (TSDC [high vs. low] × Autocorrelation [high vs. low]), entirely between subjects, and participants only worked with one scenario.

Procedures. Procedures were very similar to Experiment 1, except for the following. Participants were asked to imagine that they had chronic back pain and made an appointment with a doctor. The appointment was in 14 days, and during the initial 14 days before the appointment, they observed the pain score. On visiting the doctor, they were told to test two medicines for 14 days, one on each day, and, at the end of the 14 days, to make a decision about which medicine worked better so that it could be prescribed for the indefinite future.

Beliefs about autocorrelation were manipulated by using baseline trends that had high or low autocorrelation for all 28 days (UTW or UTW randomized functions). This meant that participants had 14 days to observe whether the baseline pain trend had high or low autocorrelation before entering the 14 days of testing. After the initial 14 days but before testing the two medicines, participants judged the amount of autocorrelation in the back pain using the same scales from Experiment 3.

Participants' beliefs about TSDC were manipulated through the simulated visit with the doctor. Participants were told that the medicines started to work in 30 min (low delay) versus 1–2 days (high delay), that they continued to work for 12 hours (low carryover) versus 1–2 days (high carryover), and that they either did not (low tolerance and sensitization) versus might start to work better or worse after repeated use (high tolerance and sensitization). All four TSDC effects were manipulated simultaneously to be either high or low. To move forward with the study, participants had to correctly answer three questions about whether the medicines exhibited TSDC effects to verify that they had read these instructions.

Results

Six participants who tried the same medicine for all 14 days were dropped from all analyses.

Did TSDC and autocorrelation beliefs influence alternation? To verify that participants were sensitive to the autocorrelation manipulation, the two autocorrelation questions were compared across the autocorrelation high versus low conditions. The two measures correlated moderately ($r = .49$, $p < .001$), and both received significantly higher scores in the autocorrelation high condition, Question 1 ($M = 5.8$ vs. 4.0), $t(196.07) = 6.71$, $p < .001$, $d = 0.95$; and Question 2 ($M = 7.0$ vs. 3.0), $t(198.96) = 12.89$, $p < .001$, Cohen's $d = 1.82$.

Figure 12 shows histograms of the number of alterations by condition. There appeared to be considerably more alternation in the TSDC low condition than the TSDC high condition; however, it was less evident whether the autocorrelation manipulation had an influence on alternation. Moreover, most participants across all conditions alternated fairly little.

A multinomial regression predicting alternation pattern (1 vs. 2–12 vs. 13) was run with autocorrelation and TSDC conditions as predictors. Increasing TSDC beliefs increased the ratio of one versus 2–12 alternations ($b = 1.35$, $SE = 0.32$, $p < .001$), and (marginally) decreased the ratio of 13 versus 2–12 alternations

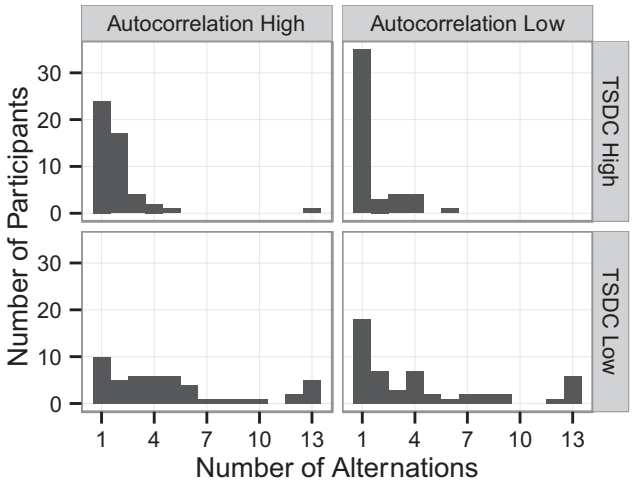


Figure 12. Histograms of the alternation in Experiment 4. TSDC = tolerance, sensitization, delay, or carryover.

($b = -1.86$, $SE = 1.06$, $p = .08$). The effect size for the total influence of TSDC was $\eta_p^2 = .09$.

Increasing beliefs about the amount of autocorrelation decreased the ratio of one versus 2–12 alternations ($b = -0.95$, $SE = 0.32$, $p = .003$). This effect can be seen primarily as a shift from one alternation (autocorrelation low) to two alternations (autocorrelation high), primarily in the TSDC high condition (see Figure 12). The autocorrelation manipulation did not influence the ratio of 2–12 versus 13 alternations ($p = .67$).[14] The effect size for the total influence of autocorrelation was $\eta_p^2 = .03$.

In sum, Experiment 4 supported the interpretation of Experiment 3 that participants choose to alternate less if they believed that TSDC effects were plausible. Though participants also alternated more when the baseline function was highly autocorrelated as opposed to not, this effect involved a small shift from one alternation to two. Even in the autocorrelation high/TSDC low condition, the condition in which alternation was strongly justified, a large majority of participants alternated less than chance. This finding supported the result from Experiment 3 that autocorrelation beliefs have a modest effect on the amount of alternation.

Sequential testing strategies. Table 3 presents bivariate correlations between the four heuristics and whether a participant would alternate or not for a given trial in the testing phase; ΔWSLS was the strongest predictor. Next, the four heuristics were entered into a multivariate logistic regression with a by-subject random intercept (a model with random slopes would not converge). Of the four heuristics, only ΔWSLS was significant (see Table 3). (A model with ΔWSLS as the only predictor and a by-subject random intercept and random slope for ΔWSLS was significant, $b = 0.57$, $SE = 0.10$, $p < .001$, $r^2 = .03$.) A follow-up regression with a random intercept testing the ΔWSLS × Autocorrelation interaction found that the effect of ΔWSLS was higher in the autocorrelation high than the autocorrelation low condition (see Table 3). This interaction effect is plotted in the Appendix and replicates Experiment 3.

Did people who alternated in autocorrelated conditions make better inferences? The same analysis from Experiment 3 was run to examine the relation between alternation and accuracy in the autocorrelation high versus the low conditions (see Table 4, Supplemental Materials 2 for figures). Though the directions of the results were as expected, the results were only marginal. The weaker results compared to prior experiments was not surprising given that this study had fewer subjects, and that many participants alternated very few times in both TSDC high conditions.

Heuristics for inferring comparative effectiveness. Both bivariate correlations and multivariate standardized regressions found that both NMH and ΔNMH predicted considerable amounts of variance in subjects' comparative efficacy judgments (see Table 5). Two-way interactions between the two heuristics and autocorrelation condition found that the effect of NMH decreased from the autocorrelation low to autocorrelation high condition, and the effect of ΔNMH increased (see Table 5).

Discussion

Experiment 4 largely replicated Experiment 3. There was only a small influence of the autocorrelation manipulation on the amount of alternation. However, participants switched toward focusing more on change scores than on raw scores for both testing the causes and inferring comparative efficacy. These findings suggested that the adaption to autocorrelation was more reactive to the experienced data than proactive to plan ahead for the autocorrelation.

In Experiment 3, the response to TSDC beliefs was somewhat unclear. In Experiment 4, there was a clear effect that participants alternated less when they believed TSDC effects to be likely, implying that the response to TSDC beliefs is proactive.

Experiment 5

Many, perhaps most, real-world information search situations do not specify a specific amount of data to be sampled. For example, instead of having exactly 14 days to test the two medicines, a patient may try the medicines until they are satisfied they have identified the better one. Do the general patterns of results extend to an information search task in which participants get to sample until they have decided which cause is best?

There are a number of reasons that a free sampling paradigm as opposed to a fixed search length could change the amount of alternation. First, when participants are told to sample for a specific amount of time (e.g., exactly 14 days), they may have a default tendency to try one medicine for 7 days and then the other for 7 days; whereas if they are given an unlimited number of samples, there is no half-way point, in which case participants might be more inclined to alternate.

Second, it is possible that the free sampling paradigm could change the extent of exploitation. On the one hand, the free sampling paradigm could potentially increase the desire to exploit; because the learner can choose how long to sample, he or she should try to get the best outcomes on each day. On the other hand,

[14] For the sake of robustness, a Poisson regression was also run to model the entire distribution rather than just the three categories of alternation. Higher tolerance, sensitization, delay, or carryover (TSDC) beliefs caused less alternation ($b = -0.97$, $SE = 0.08$, $p < .001$). Higher autocorrelation beliefs caused a marginally higher amount of alternation ($b = 0.15$, $SE = 0.08$, $p = .06$).

because participants have to decide each day to continue collecting more evidence or to terminate the search and make the final choice, the repeated choice may highlight the goal of deciding which medicine is best (exploration), leading to alternation.

A third possibility is that, through the process of trying the two medicines, participants might realize that perseverating in an autocorrelated environment does not provide definitive evidence, and transition toward alternating. With small samples, it could be easy to confuse autocorrelation for differential effectiveness, but, with longer samples, participants would see periods of time in which the pain is both high and low for both medicines, which might make them understand the need to alternate.

This experiment was not designed to distinguish all of these possible effects—the goal was to examine whether the main findings from the prior studies would hold when participants got to decide when to stop sampling.

Method

Participants. Two hundred participants (52% female) were recruited from MTurk, and 197 fully completed the study. They were paid $1, with the possibility of a bonus explained below.

Design and procedure. The design and procedures were very similar to Experiment 4, except for the following changes. First, participants could choose to test the medicines for however long they liked. Participants who reached the 50th sample were forced to make a final choice; however, they were not told about this cutoff in advance.

Second, the study design had two conditions: autocorrelation high versus low. Both conditions used the TSDC low framing from Experiment 4. Similar to Experiment 4, participants had 14 days to observe the baseline pain function before starting to test the medicines.

Third, the bonus payment was increased; participants earned 50, 40, 30, 20, 10, or 0 cents if the causal efficacy judgment was within ±0, 2, 4, 6, 8, or more than 8 points of the correct answer, respectively. The increased bonus was intended to encourage participants to focus even more on the goal of identifying the best medicine (exploration, not exploitation), and to ensure that participants had sufficient motivation not to stop the search after just a couple choices.

Results

Five participants stopped searching after only 1 day of testing, and one other participant only tried one medicine and never tried the other; their responses are not analyzed. Thus, a total of 191 participants were included in the analysis.

Did autocorrelation beliefs influence alternation? Replicating Experiment 4, the autocorrelation manipulation worked. Mean autocorrelation ratings were higher for both measures in the high condition ($M = 6.0$, $M = 6.9$) than for the respective measures in the low condition ($M = 3.9$, $M = 2.8$), ts(189) > 9.23, ps $< .001$, Cohen's ds > 1.34. The next question was whether participants modified their search strategies in response to the autocorrelation; the total number of choices, total number of alternations, and percentage of alternations were all assessed.

Figure 13 shows a histogram of the number of choices made before terminating the search, separated by the two conditions. The mean for the autocorrelation high condition was 18.5 and the mean for the autocorrelation low condition was 16.9. Only four participants were cut off at 50 choices. A Poisson regression was used to test whether there was a difference in the two conditions. Participants in the autocorrelation high condition tested the medicines a little longer ($b = 0.09$, $SE = 0.03$, $p = .007$, $r^2 = .006$).

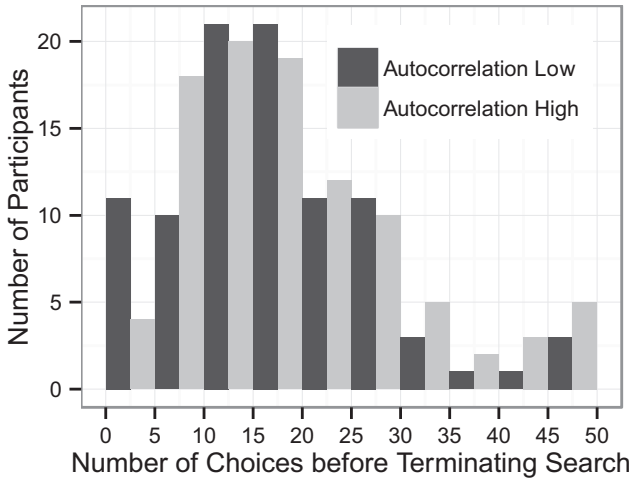


Figure 13. Histogram of the number of choices before search termination in Experiment 5.

Figure 14 shows a histogram of the total number of alternations per participant. Most participants switched between the two medicines less than five times, often only once. A Poisson regression revealed more alternations in the autocorrelation high condition ($b = 0.18$, $SE = 0.06$, $p = .002$, $r^2 = .009$).

Figure 15 shows a histogram of the percentage of alternations per subject. About 15% of participants alternated at every opportunity, but 68% of participants alternated less than 50% of the time. Because the distribution was bimodal, a logistic regression was used to determine whether there was a different ratio of alternating at every opportunity (1) versus less than every opportunity (< 1) across the two conditions; this effect was not significant ($b = 0.22$, $SE = 0.41$, $p = .57$). I then looked to see whether there was an effect of the autocorrelation manipulation just within participants who alternated less than 100% of the time. A Wilcoxon's test for independent samples was not significant ($p = .13$, $r^2 = .008$), suggesting again that if there was an influence of the autocorrelation manipulation on information search, it was small.

Sequential testing strategies. Table 3 presents both bivariate and multivariate analyses of the relation between the four heuristics and alternation. Due to convergence difficulties, the multivariate analysis did not have random effects on the slopes but had a random effect on the intercept. ΔWSLS was the only significant predictor. A follow-up test with a random intercept found that subjects' testing decisions were more sensitive to ΔWSLS in the autocorrelation high than the autocorrelation low condition (see the Appendix for a figure). This interaction replicated the same findings in Experiments 3 and 4. The fact that the effect was weaker in this experiment is probably because the rate of alternation was so low across both conditions.

Changes in alternation across time. A new question for this experiment is whether learners increased or decreased the amount of alternation across time, especially within the autocorrelated

condition. It is possible that after multiple days of trying one medicine, they could come to realize the importance of alternating.

Analyzing the amount of alternation across time in a free search paradigm is complicated. In prior research, participants who alternated more frequently tended to terminate the search earlier (Hills & Hertwig, 2012; Rakow, Demes, & Newell, 2008), which can make it appear as if there is a trend toward increased perseveration across time. Indeed, a correlation between the number of samples before terminating the search and the percentage of alternations from the number of possible alternations was negative ($r = -.37$, $p < .001$). Different methods of accounting for this confound can lead to different conclusions (Gonzalez & Dutt, 2012; Hills & Hertwig, 2012). Thus, a variety of analytical approaches were used to answer this question.

Hills and Hertwig (2012) recommended comparing the percentage of alternations within the first 25% and last 25% of opportunities to alternate. This analysis, using only participants who had at least four opportunities to alternate, revealed no difference within the first and last quarters, $t(179) < 1$ ($p = .60$). Additionally, the difference between the beginning and end was not influenced by the autocorrelation manipulation, $t(178) < 1$ ($p = .91$).

Gonzalez and Dutt (2012) looked within subsets of participants who had at least 6, 10, and 18 opportunities to alternate (7, 11, and 19 choices before stopping sampling), and then tested whether the amount of alternation differed within the first half and the second half of the n samples. Using this approach, there also were no changes in alternation for any of the subsets; p values ranged from 0.29–1.00 for the autocorrelation high condition and from .46–.69 for the autocorrelation low condition.[15]

This finding of no systematic changes in the amount of alternation over time fits with the use of ΔWSLS compared with NMH and ΔNMH. ΔWSLS is largely driven by the increasing versus decreasing periods in the baseline trend. Consequently, there would be little systematic change in the amount of alternation at the beginning versus the end of the learning sequence. In contrast, NMH and ΔNMH are built to exploit based on all prior experiences rather than just the most recent experience and, consequently, they predict a shift toward less alternation over time.

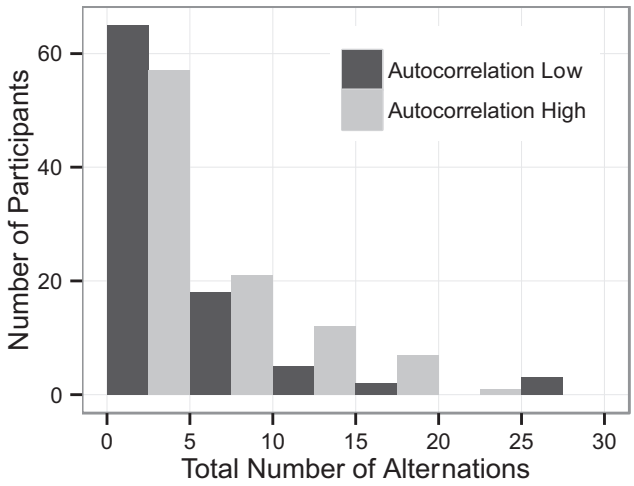


Figure 14. Histogram of the total number of alternations per participant in Experiment 5.

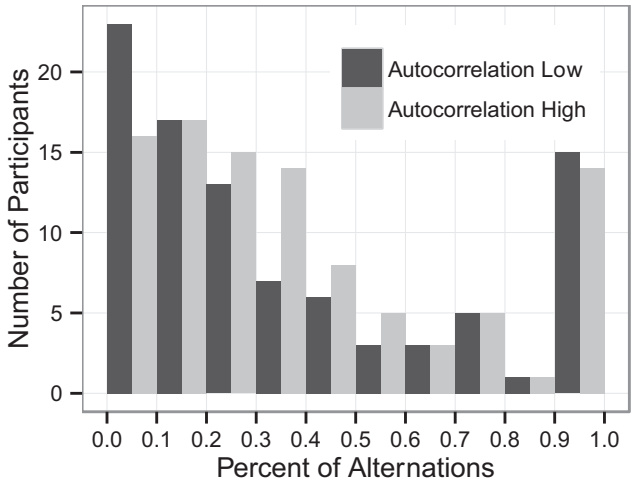


Figure 15. Histogram of the percentage of alternations in Experiment 5.

Did people who alternated in autocorrelated conditions make better inferences? I examined the relations between the total number of samples before termination, total number of alternations, and percentage of alternations on accuracy (see Supplemental Materials 2 for plots). The total number of samples did not correlate with log error either for the autocorrelation high condition ($p = .64$) or for the autocorrelation low condition ($p = .29$).

Replicating the previous experiments, the number of alternations and percentage of alternations were associated with less absolute logged error and better ability to choose the better cause in the autocorrelation high condition, but not in the autocorrelation low condition, and the interactions between these two measures and alternation condition were significant (see Table 4). There was one difference from the prior experiments: for the first time, a higher percentage of alternations was associated with higher error in the autocorrelation low condition. It is possible that perseveration is cognitively beneficial (cf. Hills & Hertwig, 2012), or this result may have been complicated by the fact that higher rates of alternation were associated with testing fewer samples.

Heuristics for inferring comparative effectiveness. Table 5 reports the bivariate, multivariate, and two-way interactions between NMH and ΔNMH and autocorrelation on comparative effectiveness. The findings replicated the previous experiments that subjects focused more on change scores for the autocorrelation condition and more on raw scores for the autocorrelation low condition.

Discussion

Experiment 5, in which subjects could decide when to stop the search, largely replicated the previous studies. First, the overall rate of alternations was still fairly low. Second, there were only fairly small

[15] I also conducted the same analysis with the last n opportunities to alternate. This analysis also revealed no effects; p values ranged from .09−.44 for the autocorrelation high condition and from 0.38−1.00 for the autocorrelation low condition.

changes in the overall amount of alternations based on the amount of autocorrelation. Third, there was a shift in participants' search strategies toward using change-based strategies for deciding which cause to try at each opportunity and for assessing comparative accuracy when autocorrelation was high compared to low.

General Discussion

Five experiments examined how individuals test which of two causes is better. When they believed that tolerance, sensitization, delay, and carryover (TSDC) effects were plausible, they tried each cause longer. The influence of beliefs about autocorrelation was more subtle. One hypothesis was that participants would increase the number of alternations to adapt to alternation; however, this effect was small, when present, despite that alternation improved accuracy in autocorrelated environments. Furthermore, the low rates of alternation (almost always $< 50\%$) cannot be explained by a general cognitive benefit of perseveration; when autocorrelation was low, forcing participants to alternate or perseverate had no influence on the accuracy of identifying the better cause. However, participants adapted to the autocorrelation in two other ways.

Experiments 3–5 (see the Appendix) found that participants adapted to the autocorrelation in the baseline function by using the ΔWSLS strategy more when autocorrelation was high. ΔWSLS involves staying with the same cause if it results in an advantageous change in the outcome from the previous period, otherwise switching. Simulations showed that, in autocorrelated environments, ΔWSLS permitted a high degree of accuracy in assessing comparative efficacy while simultaneously permitting the learner to exploit the better cause to a certain amount. Experiments 2–5 also found that participants adapted by assessing comparative efficacy using change scores more in autocorrelated environments and raw scores more in environments without autocorrelation.

In summary, the change in the rate of alternation based on TSDC beliefs was most likely a product of proactive planning. In contrast, the adaptations to autocorrelation, which mainly occurred when assessing comparative efficacy and to a lesser extent in the sequential search, were more consistent with a reactive shift in strategy based on the experienced data.

For consistency across findings, I adopted percent variance explained measures of effect size (r^2 and η_p^2). Cohen's d interpretations of small, medium, and large (.2, .5, > .8) translate into variance explained of roughly .01, .06, and .14. This means that the benefits of alternation in the autocorrelation condition were mainly medium to large, though the Alternation × Autocorrelation interactions ranged greatly. In terms of adapting to autocorrelation, the increases of alternation (when significant) were small, the increases in the use of ΔWSLS as a sequential strategy were small, and the increases in the use of ΔNMH due to autocorrelation for the final calculation of comparative efficacy were medium (higher in Experiment 3–5 when participants chose their samples). The influence of TSDC beliefs on the amount of alternation was medium.

Metacognitive Understanding of Planning for Autocorrelation

To better understand participants' reasoning, I added questions to the end of the studies (see Supplemental Materials 1 for all questions). After the comparative efficacy judgment, participants also rated their confidence. Across Experiments 2–5, there was no consistent pattern and, for the most part, the confidence judgments were unrelated to the amount of alternation. There were slightly *positive* (sometimes significantly positive) correlations between the confidence judgments and absolute logged error in the comparative efficacy judgment. Participants may have confused increasing (or decreasing) patterns in the baseline function to the drugs and thought that they worked very poorly (or very well) with high confidence.

In Experiment 5, participants rated the extent to which they alternated due to autocorrelation: "I switched back and forth between the medicines because I was worried that there could be periods in which my pain is naturally high or naturally low. So I wanted to switch off between the medicines to see if the medicines made a difference or if the pain was just changing by itself over time." Participants endorsed this question equally across the high versus the low autocorrelation conditions, $t(184) = 1.34$ ($p = .17$), even though they knew the degree of autocorrelation before starting the search task. These findings suggest that participants had minimal foresight to increase the amount of alternation in the face of autocorrelation.

Exploitation Versus Positive Testing

Despite the fact that participants were only paid for their final choice of which option was better (exploration), the fact that the sequential strategies predicted subjects' testing choices is evidence that participants engaged in exploitation and/or positive testing. The effect of exploitation and/or positive testing can be quantified in the following way: after ignoring the 47% of scenarios in which participants tested both causes the same number of times and scenarios in which participants concluded that the causes were equally effective, 81% of the time participants tested the cause that they eventually chose as the better cause more frequently than the cause that they eventually selected as the worse cause.

It is difficult to empirically differentiate between exploitation and positive testing because it predicts the same behavior. However, at the end of Experiments 3–5, participants were asked to rate the extent that they used these two strategies (see the Method section in Experiment 3 for the wording). There was a consistent pattern of slightly higher ratings for positive testing than exploiting: Experiment 3 ($p = .06$, $d = 0.10$), Experiment 4 ($p < .001$, $d = 0.25$), and Experiment 5 ($p = .002$, $d = 0.22$). There are two other reasons to believe that exploitation did not play an outsized role in this study. First, Gonzalez and Dutt (2012) analyzed two datasets involving a task that was supposed to involve only exploration, just like the current experiments, and they also found a correlation between perseverating on one option and choosing that option when making the final choice. Second, there was no transition toward more perseveration across time in Experiment 5, which is a typical pattern in tasks with an exploration−exploitation tradeoff (Hills & Hertwig, 2012).

It is not entirely clear how exploitation and positive testing would play out in the real world. One hypothesis is that these motivations would become even larger when the outcome variable has more significant consequences, which could result in an even stronger use of sequential as opposed to preplanned testing strategies. Furthermore, an increase in these motivations could result in a shift from ΔWSLS and toward ΔNMH, which shifts the balance more toward exploitation and away from exploration (see Table 2).

Attending to Change Scores in Time Series Data

One novel finding was that participants' judgments of comparative efficacy were predicted not only by NMH, but also by the difference in the average change scores associated with the two causes (ΔNMH). This strategy makes sense from a time series analysis perspective, because taking a difference score helps to counter the effects of nonstationarity. This use of change scores also fits into a larger body of work that people use change scores for other types of causal inference, such as for inferring the direction of a causal relation (Rottman & Keil, 2012; Rottman, Kominsky, & Keil, 2014). Given that everyday variables are positively autocorrelated, focusing on changes may be a useful strategy for causal reasoning.

Future Directions

The current findings raise three important questions for future research. First, the current research does not answer whether individuals use multiple search strategies or just one. Second, though the current studies identified some factors that influence the causal testing process, there is considerable variance left to be explained. In particular, Experiment 3, which tested 15 different cover stories, found two stories in which the amount of alternation was high relative to all the other conditions. Both involved situations in which TSDC effects were not plausible and autocorrelation was high, which may partially explain the effect; but other analogous cover stories had more moderate rates of alternation. This raises the possibility that domain-specific assumptions may play important roles in causal testing strategies.

Third, future research should attempt to develop an optimal exploration strategy that learns about autocorrelation in the baseline function and TSDC effects. I suspect that this task will be very challenging because (a) the possible baseline functions are unlimited, (b) TSDC effects can be very complicated and may require something like a pharmacodynamics model, and (c) it may be hard to distinguish nonstationary baseline functions from TSDC effects, especially from limited numbers of samples. Despite these challenges, such a model may be useful for guiding decision-makers searching for the best cause. Such a model may also be able to explain overlooked patterns in individuals' search behaviors and learning difficulties such as discriminating TSDC effects from changes in the baseline function.

Conclusion

The current experiments examined how individuals test which of two causes produces a better outcome. This goal is very complicated in that the number of possible testing strategies is large, and the utility of various testing strategies depends on both the baseline function and whether the causes exhibit TSDC effects. Though many participants could have performed better had they alternated more in autocorrelated environments, they adapted their testing strategies to different environments in other ways. When deciding which cause worked best, participants adaptively shifted between focusing on raw outcomes to changes in outcomes for autocorrelated environments. More broadly, it will be important to test whether these findings generalize to real-world settings such as doctors and patients trying new medications, or managers or policymakers deciding whether a new intervention such as an advertising campaign or policy is working. Understanding habits of testing strategies in real-world domains may identify cases in which causal learning can be improved by helping decision-makers understand how to adapt their testing strategies to the particular environment.

References

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12,* 199–210. http://dx.doi.org/10.1901/jaba.1979.12-199

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology, 53,* 155–167. http://dx.doi.org/10.1016/j.jmp.2008.05.006

Bose, M., & Dey, A. (2009). *Optimal crossover designs*. London, England: World Scientific.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 708–731. http://dx.doi.org/10.1037/xlm0000061

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8,* 269–295. http://dx.doi.org/10.1080/13546780244000060

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. http://dx.doi.org/10.1037/0033-295X.104.2.367

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology, 79,* 102–133. http://dx.doi.org/10.1016/j.cogpsych.2015.02.004

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. http://dx.doi.org/10.1038/nature04766

Gonzalez, C., & Dutt, V. (2012). Refuting data aggregation arguments and how the instance-based learning model stands criticism: A reply to Hills and Hertwig (2012). *Psychological Review, 119,* 893–898. http://dx.doi.org/10.1037/a0029445

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. http://dx.doi.org/10.1016/j.cogpsych.2005.05.004

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7,* 464–481. http://dx.doi.org/10.1177/1745691612454304

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137. http://dx.doi.org/10.3758/BF03194330

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518. http://dx.doi.org/10.1002/bdm.598

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 209–236). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0010

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115,* 225–237. http://dx.doi.org/10.1016/j.cognition.2009.12.009

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience. Do our patterns of sampling foreshadow our decisions? *Psychological Science, 21,* 1787–1792. http://dx.doi.org/10.1177/0956797610387443

Hills, T. T., & Hertwig, R. (2012). Two distinct exploratory behaviors in decisions from experience: Comment on Gonzalez and Dutt (2011). *Psychological Review, 119,* 888–892. http://doi.org/10.1037/a0028004

Jones, B., & Kenward, M. (2003). *Design and analysis of cross-over trials*. Boca Raton, FL: Chapman & Hall/CRC.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856–876. http://dx.doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 451–460. http://dx.doi.org/10.1037/0278-7393.32.3.451

Laska, E., Meisner, M., & Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics, 39,* 1087–1091. http://dx.doi.org/10.2307/2531343

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General, 143,* 94–122. http://dx.doi.org/10.1037/a0032108

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making, 7,* 119–148.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology, 22,* 408–412. http://dx.doi.org/10.1016/0022-0965(76)90104-1

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112,* 979–999. http://dx.doi.org/10.1037/0033-295X.112.4.979

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science, 21,* 960–969. http://dx.doi.org/10.1177/0956797610372637

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106,* 168–179. http://dx.doi.org/10.1016/j.obhdp.2008.02.001

Ratikowsky, D., Evans, M., & Alldredge, J. (1993). *Cross-over experiments: Design, analysis, and application*. New York, NY: Marcel Dekker.

Rottman, B. M., & Ahn, W. K. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin & Review, 16,* 1043–1049. http://dx.doi.org/10.3758/PBR.16.6.1043

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology, 64,* 93–125. http://dx.doi.org/10.1016/j.cogpsych.2011.10.003

Rottman, B. M., Kominsky, J. F., & Keil, F. C. (2014). Children use temporal cues to learn causal directionality. *Cognitive Science, 38,* 489–513. http://dx.doi.org/10.1111/cogs.12070

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701. http://dx.doi.org/10.1037/h0037350

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5,* 472–480.

Senn, S. (1993). *Cross-over trials in clinical research*. Chichester, England: Wiley.

Shumway, R., & Stoffer, D. (2011). *Time series analysis and its applications: With R examples*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-7865-3

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. *Statistical Science, 5,* 465–472.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53,* 168–179. http://dx.doi.org/10.1016/j.jmp.2008.11.002

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489. http://dx.doi.org/10.1207/s15516709cog2703_6

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review, 20,* 364–371. http://dx.doi.org/10.3758/s13423-012-0324-9

Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W. K. Estes. *Journal of Mathematical Psychology, 59,* 41–49. http://dx.doi.org/10.1016/j.jmp.2013.10.001

Yi, S. K. M., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving, 2,* 81–102. http://dx.doi.org/10.7771/1932-6246.1060

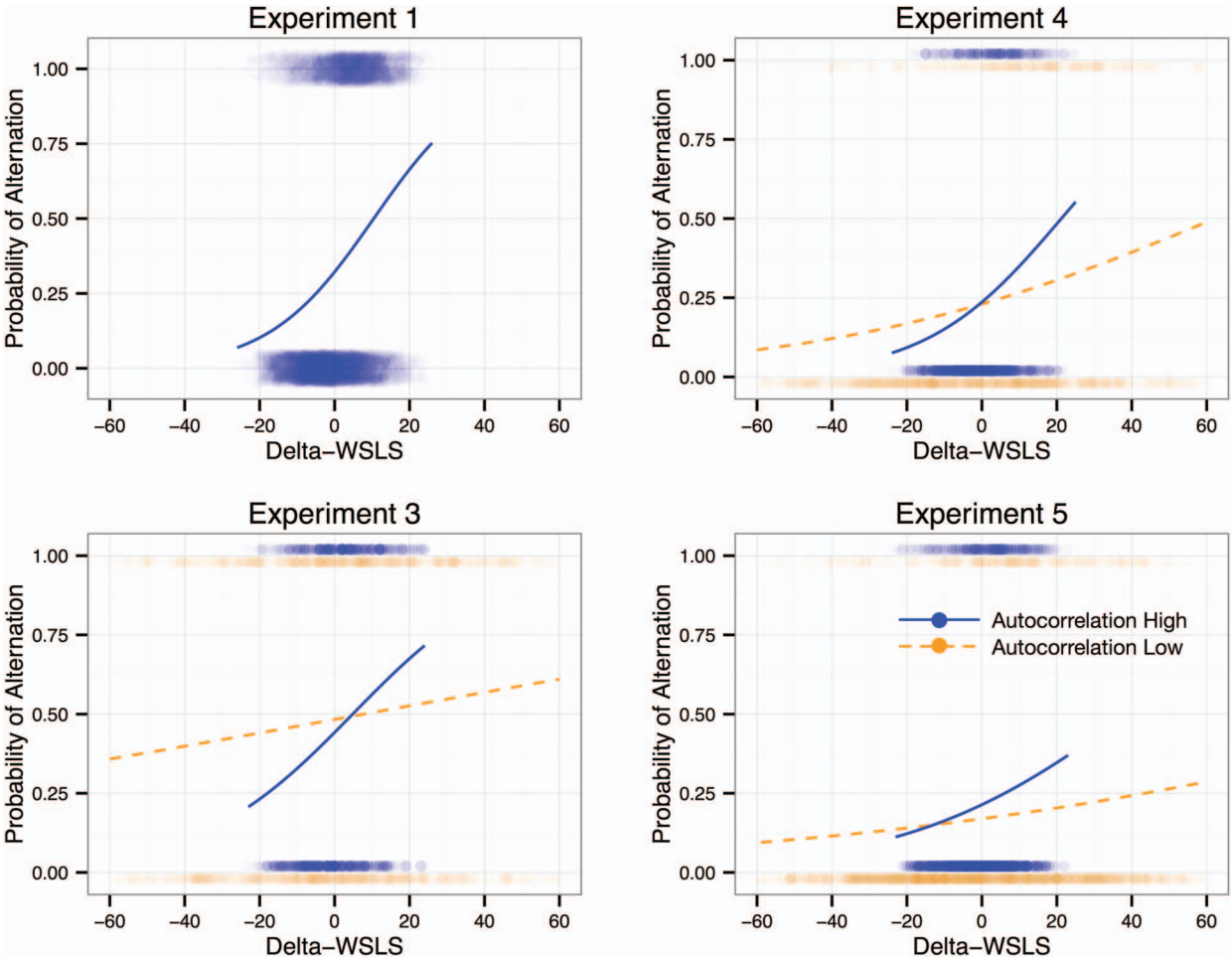
(*Appendix follows*)

Appendix

The Relation Between the ΔWSLS and the Probability of Alternation

These figures show the relation between the change-score win−stay lose−shift (ΔWSLS) and the probability of alternating on the next trial. The actual choices to alternate are plotted at the top (switch to other cause) and bottom (stay with current cause) with transparency due to the large number of choices. The curves are logistic regression lines, which can be interpreted as the average probability of alternating given a particular change in the outcome (ΔWSLS). Experiment 2 is missing because, in that study, participants were not allowed to choose when to alternate.

In the autocorrelation high conditions, the variance of ΔWSLS was lower than in the autocorrelation low conditions. The interactions show participants' choices were more sensitive to ΔWSLS in the autocorrelation high conditions. A larger increase in the outcome was needed to prompt the same probability of a change in the autocorrelation low conditions than in the autocorrelation high conditions.



Note. See the online article for the color version of this figure.

Received March 7, 2015
Revision received December 9, 2015
Accepted December 15, 2015 ■