BARBARA A. SPELLMAN, Associate Professor, Department of Psychology, University of Virginia

ALEXANDRA P. KINCANNON, Ph.D., Department of Psychology, University of Virginia

STEPHEN J. STOSE, Ph.D. Candidate, Department of Psychology, University of Virginia

Barbara A. Spellman
Department of Psychology
University of Virginia
P.O. Box 400400
Charlottesville, VA  22904-4400
USA

TEL: (434) 982-5591
EMAIL: spellman@virginia.edu

Final draft.  To appear as:

THE RELATION BETWEEN COUNTERFACTUAL AND CAUSAL REASONING

Barbara A. Spellman, Alexandra P. Kincannon, and Stephen J. Stose

INTRODUCTION

Among the vast array of thinking skills that humans possess are the abilities to engage in counterfactual and causal reasoning. Counterfactual reasoning allows us to imagine something in the world being other than it actually was or is (i.e., counter-to-fact); we can then imagine, or mentally simulate, the world continuing to unfold in a direction other than the direction it has actually taken. This ability allows us to torment ourselves with regret ("If only I hadn't gone for a drive that night...") or to create our own personalized version of It’s a Wonderful Life. It also allows us to plan for the future and to learn from our mistakes. Whereas counterfactual reasoning is about possibility, causal reasoning is about reality. Finding causes is at the heart of our scientific endeavors; assessing causality is essential for meting out justice in our legal system. As many researchers are fond of noting, discovering causal relations allows us to understand, explain, predict, and control our world.

In the past researchers have posited various relations between the two types of reasoning; for example, that counterfactual reasoning was at the heart of causal reasoning or that they were fundamentally identical (for history see Spellman and Mandel 1999, 2003). Currently, however, researchers seem to agree that causal and counterfactual reasoning ask different questions and serve different functions (Mandel 2003c; Mandel and Lehman 1996; also Roese 1994). In counterfactual reasoning we focus on prevention; we ruminate about counterfactuals that might
have prevented an outcome (e.g., “if only I hadn’t gone for that drive”). In causal reasoning we focus on how an event actually occurred (e.g., a man ran the red light smashing into my car).

Given the history of the investigation, researchers must acknowledge that despite their different foci, counterfactual and causal reasoning are related. Our goal in this chapter is to offer answers to the questions: what is that relation and why does it exist?

One suggestion, gleaned from the early literature, might be this: Events people select as “causes” must be earlier events (“antecedents”) that pass the but for test for later outcomes (“consequences”). The but for test of causality relies on counterfactual reasoning as illustrated by the philosopher Mackie (quotation from Lipe 1991: 457-58):

Mackie (1974) argued that our concept of causation is intimately tied to such counterfactual questions. He argued that when we are able to imagine or observe instances of the effect (the shattering of the goblet) without the proposed cause (the high note), causality is not inferred. However, when we are unable to imagine such alternative situations, the proposed causal link remains intact. Hart and Honoré (1959), cited in Mackie (1974), stated that the lawyer approaches causal statements in the following way, "When it is suggested that A is the cause of B he is apt to ask as the first question, would B have happened without A?" (Mackie, 1974, p. 121).

Note the two steps in inferring causality: first, one must be able to imagine that A (the
antecedent) can be changed or mutated; and second, such a change must "undo" B (the consequence). Thus, another way to state the proposed relation between causal and counterfactual reasoning would to be to say that for an antecedent to be a called a “cause" of a consequence, if the antecedent were changed the consequence would be changed. Typically, when participants are asked to select a cause of an outcome, they will select a cause but for which the outcome would not have occurred (e.g., the man running the red light is a cause; if he hadn't done so then there would have been no accident).

However, simple as the example of the but for test seems, and as important as it is in the legal system (Hart and Honoré 1985; see Spellman and Kincannon 2001), that test does not work for all causes. The obvious exceptions are situations involving multiple sufficient causes: that is, when two (or more) causes act simultaneously or sequentially and either cause alone would be enough to cause the outcome.

Simultaneous multiple sufficient causes are illustrated by an example from Spellman and Kincannon (2001). Participants read the following story:

Reed hates Smith and wants to kill him. West also hates Smith (for an entirely different reason) and also wants to kill him. One day Reed shoots Smith in the head. At the exact same instant, West shoots Smith in the heart. Smith dies. The coroner says that either shot alone would have been enough to kill Smith.

Participants realized that changing Reed or West’s actions alone would not change the outcome
(death) yet they attributed full causality to both Reed and West individually and sentenced them each to maximum jail time.

Sequential multiple sufficient causes occur when, for example, A gives X a lethal dose of poison, but before it can take effect, B shoots X dead (see Katz 1987 for the classic legal examples). In such cases, participants tend to attribute causality to B, even though changing B’s actions would not change the fact of X’s death (Greene and Darley 1998; Mandel 2003c). Thus, cases of multiple sufficient causes show that but for causality does not capture the psychological relation between counterfactual and causal reasoning.

Instead, the suggestion we offer here is that counterfactual and causal reasoning are similar (a) because they rely on the same underlying information and (b) because the counterfactuals people consider provide input into their computations of causality.

**OUR THEORY OF THE RELATION BETWEEN COUNTERFACTUAL AND CAUSAL REASONING**

Our theory of the relation between counterfactual and causal reasoning (or, more accurately, between mutability and causality judgments) can best be explained by reference to one figure and one equation. The figure illustrates the way that mutability and causality judgments are based on similar information; the equation details how specific causality judgments might rely on specific counterfactual information.

Before describing the details of our theory, however, we should explicate what kind of
judgments it makes predictions about. In experiments, mutability and causality judgments are usually elicited after participants read a scenario. When making causality judgments, participants are typically asked to either list causes of the outcome, rate events that are on a provided list of potential causes, or both. Counterfactual judgments (mutability judgments) are typically solicited in one of two ways. Participants may be asked how a character in the story would complete a sentence beginning with “if only” (as in Kahneman and Tversky 1982; Mandel and Lehman 1996). Alternatively, they may be asked to list some number of ways in which the story could be changed (i.e., “mutated”) so that the outcome would be different or "undone" (e.g., Girotto et al. 1991; Wells and Gavanski 1989; Wells et al. 1987).

The Information

The essence of our figure is this: Mutability and causality judgments rely on similar underlying information but do not rely directly on each other. (Note that this does not preclude one type of judgment from influencing the other indirectly, as explained below.) The relevant information is: (a) the availability of alternatives to the cause and effect and (b) pre-existing knowledge - in particular, pre-existing causal knowledge.

Pre-existing causal knowledge is necessary for both kinds of judgments. To make a mutability judgment, the reasoner must know whether changing an antecedent would change the consequence; such a judgment requires causal knowledge. To make a causality judgment, the reasoner must have either knowledge of covariation, beliefs about causal mechanisms, or both.
Alternatives to the cause and effect may become more available through both external and internal cues. External availability cues are provided by the experimenter. For example, the order in which a story is presented can affect the availability of alternatives (Byrne et al. 2000; Spellman 1997, 2003) as can the way in which questions about the events are phrased (Mandel and Lehman 1996). Internal availability cues are provided by the participant. For example, when a participant judges both mutability and causality in a single experiment, answers considered during the earlier judgment are more available for the later judgment (Spellman, 2003).

Consequences of such order effects are described below.

**The Equation: Causality Judgments for Single-Occurrence Events**

Following Spellman’s (1997) SPA model (“Spellman probability-updating account” coined by Mandel 2003c), we assume that a causality judgment about a person or event is a function of how much that person or event increases the probability of the outcome above its previous probability.\(^1\)

Below is a simple equation depicting that function.

\[
C \approx p(\text{after}) - p(\text{before})
\]

Equation 1 says that causality (C) is a function of the probability of the outcome occurring estimated *after* the target event has occurred ("probability after") minus the probability of the outcome occurring estimated *before* the target has occurred ("probability before"). Note that
when there are a series of events that might be viewed as causal, each estimate can be inserted in sequence. SPA suggests that if people are asked to identify a single cause of an outcome, they will choose the cause that maximizes $C$ in Equation 1.

The probabilities in Equation 1 can be “unpacked” by expressing each as a probability divided by 1 and then expanding 1 into two complementary components. (Note that this kind of unpacking is common in psychological functions, e.g., Bayes’ Theorem.)

\[
(2) \quad C \approx \frac{p(O_{after})}{p(O_{after}) + p(\sim O_{after})} - \frac{p(O_{before})}{p(O_{before}) + p(\sim O_{before})}
\]

Given that the events in question are one-time events, how can people estimate the probability that an outcome would happen? One way would be to consider all of the ways that the world could unfold, and sum up, for each way, the product of the likelihood that that “way” would happen \([p(\text{way})]\) and the probability of the outcome given that “way” \([p(O|\text{way})]\).

\[
(3) \quad C \approx \frac{\sum_{\text{ways}} p(\text{way}) \times p(O_{after}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(O_{after}|\text{way})} - \frac{\sum_{\text{ways}} p(\text{way}) \times p(\sim O_{after}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(\sim O_{after}|\text{way})}
\]

\[
= \frac{\sum_{\text{ways}} p(\text{way}) \times p(O_{before}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(O_{before}|\text{way})} + \frac{\sum_{\text{ways}} p(\text{way}) \times p(\sim O_{before}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(\sim O_{before}|\text{way})}
\]

We can use these equations, together with the figure, to explain several phenomena involving counterfactual and causal reasoning.
Defining “The Outcome”

An important question raised by Mandel (2003c; see also Spellman and Kincannon 2001; Strassfeld 1992) is what should count as the “outcome” in these situations. For example, which was the outcome of the Reed and West story: Smith's death?; Smith's death by gunshot?; Smith's death by two gunshots?

In Mandel’s (2003c) Experiment 1, which involved sequential multiple sufficient causes, participants read about a career criminal (Mr Wallace) who is poisoned by Assassin 1, but before the poison can take effect, his car is intentionally run off the road by Assassin 2. Wallace dies when his car explodes. Participants were asked to list and then rate up to four factors that caused Wallace’s death. They were also asked to estimate the probability of Wallace dying “given” neither, each, and both of the assassination attempts. Mandel argues that according to SPA (a) participants should view the poison as most causal (because it most increases the probability of death) and (b) causal ratings should be correlated with the change in estimated probability of death attributable to each assassination attempt. Neither of those results was found. Rather, consistent with Mandel’s Judgment Dissociation Theory, participants rated the crash as most causal of Wallace’s death but assessed the administration of poison as making the biggest change to the probability of death. Mandel proposed the “actuality principle:” when selecting the most causal event, participants choose a sufficient condition that plays a direct role in generating the actual specific outcome not the event that most increases the probability of that general type of outcome.

We don’t have a problem with Mandel’s data and we like (and are not surprised by) his point;
however, we disagree with the way he applied his results to our theory. When participants are asked what caused Wallace's death, we agree that most will interpret the question as what caused him to die *in this particular manner* (i.e., car explosion). We know, in fact, that in cases of sequential multiple sufficient causes, the law (which here is designed to track human intuition; Hart and Honoré 1985) will call Assassin 2 guilty of murder whereas Assassin 1 is only guilty of attempted murder. Why? Because Wallace died from the explosion, not from the poison. We also know (Spellman 2003) that causal ratings are correlated with probability ratings only in *hindsight* - that is, only when the rater knows what *actually* caused the specific outcome. For example, “driver inattention” could be rated highly as something that might cause an accident in general, but it might not have changed the probability of a specific accident at hand.

Mandel's dependent variable question, asking for the probability of Wallace dying (generally), thus creates a problem: it calls for an interpretation. Is it asking about the probability of him ever dying? (in which case, the answer is 100 percent). About the probability of him dying in some manner having to do with the information in the story? About the probability of him dying in the exact way he actually died? We believe that different participants in different conditions interpret the question in different ways. We also believe that if the question specified the outcome, so that the interpretation was consistent across conditions (i.e., death in an off-road car explosion), the probability-change data would be correlated with the causality data, just as SPA predicts.

**THE PHENOMENA**

Thus, with a properly specified outcome, we can use the equations, together with the figure, to explain several phenomena involving counterfactual and causal reasoning.
Phenomenon 1: Effects of the Number of Counterfactual Alternatives

One prediction that Equation 2 makes is that the more alternative choices there are (some of which would lead to an alternative outcome), the more causality will be attributed to an event that causes the actual outcome to occur. That prediction might be viewed as a generalization of the well-known Wells and Gavanski (1989) wine experiment. In that experiment, participants read about a woman who was taken out to dinner by her boss. The boss orders for both of them, but the dish he orders contains an ingredient - wine - to which the woman is allergic. She eats the dish, gets sick, and dies. In both conditions of the experiment the boss had considered ordering something else: in one condition (one-wine) the alternative dish did not contain the fatal wine, in the other condition (two-wine) the alternative dish also contained wine. Participants rated his choice as more causal of the employee’s death in the one-wine condition (i.e., when he chose between one dish with it and one dish without it). Wells and Gavanski argued that these results show that counterfactual reasoning affects causal judgments because in the one-wine condition, in which there was a counterfactual alternative to the boss's decision that would have undone the outcome, the boss was viewed as more causal; therefore it must be that the availability of the non-fatal alternative dish affected that causal judgment.²

According to our equations, increasing the number of alternatives that would not lead to the outcome [i.e., increasing \( p(\neg \text{Obefore}) \)] would decrease the overall "probability before" and thus should increase the causality of an event that does lead to the outcome. The equations predict this result not only for experiments in which there is a 50 per cent (one-wine condition) versus a 100 per cent (two-wine condition) chance of choosing a bad option, but also for conditions with
varying numbers of choices.

We set out to demonstrate this predicted effect with a replication and extension of Wells and Gavanski (1989). In our modernized variant (Spellman and Meyers 2003), participants read about a seafood restaurant in which either one-of-five, three-of-five, or all five-of-five dishes that one person considered ordering for another (while the other was busy and at the other's request) contained a “spoiled” ingredient (mussels). A dish containing mussels is ordered and the other person gets food poisoning. Participants were asked how strongly they (dis)agreed with the statement that the decision to order the mussel dish caused the other person's illness (-5 = totally disagree; 0 = neutral; 5 = totally agree). Following Equation 2, participants rated the decision as most causal in the condition in which only one of the five dishes had the bad ingredient ($M = 1.0$); less causal when three dishes had the ingredient ($M = 0.7$); and even less causal when all five dishes had the ingredient ($M = -0.3$).

Thus, causal ratings not only increase when one outcome-undoing alternative is available (as in Wells and Gavanski 1989), but they also change in predictable ways depending on the number and type of alternatives considered and where those alternatives fit into the equation.

**Phenomenon 2: Order Effects when Making both Counterfactual and Causal Judgments**

Several experiments have demonstrated that judging mutability before causality will affect that later causality judgment; a few have demonstrated the opposite (see examples below). Our proposal can explain the discrepancy between these results. One key idea is that answers considered when making the first judgment become available for the subsequent judgment - that
is, that there are now internal cues to availability. The other key idea is that the existence of outcome-undoing counterfactuals affects the probability information used to judge causality.

**Mutability before Causality: The If-Only Effect**

The “if only” effect refers to the finding that when participants are asked to imagine ways in which an actor in a story might have done something different so that the outcome would not have occurred, causal ratings for those actors increase (Branscombe et al. 1996; McCloy and Byrne 2002). For example, Branscombe et al. (1996) had participants read a story about a date rape and then listen to a mock lawyer’s closing argument suggesting possible mutations to the story. If the argument mutated the defendant’s actions so that the rape would be undone, the rapist was assigned more fault (cause, blame, and responsibility) than if his actions were mutated but the rape still would have occurred. Similarly, if mutating the victim’s actions would undo the rape, she was assigned more fault than if her actions were mutated but the rape still would have occurred (see also McCloy and Byrne 2002; Spellman and Kincannon 2001).

In addition to hearing other people's suggestions for undoing the outcome, generating one's own undoing mutations can affect later causal judgments. In another experiment by Branscombe et al. (1996), some participants read the same date rape story as described above and were asked to write down a change to the victim's actions in the story. Participants who changed the victim's action such that the rape would be undone later rated the victim as more at fault than those whose changes would not have undone the rape. Similarly, Wells and Gavanski (1989) found overall that generating undoing mutations first increased later causal ratings but rating causes first did not affect later mutations.
Why should judging mutability before causality have this effect? Imagining alternatives that undo the outcome will create in the participants’ minds more ways in which the outcome could have been avoided. The effect on Equation 2 is that such imagining will increase the denominator [by increasing \( p(\neg O_{before}) \)] but not affect the numerator for the “probability before.” Thus, the value of that fraction will decrease and the value of the “probability after” minus the “probability before” will increase, resulting in more causality being attributed.

**Mutability before Causality: The Even-If Effect**

The “even if” effect is the opposite of the "if only" effect: when participants are asked to imagine ways in which an actor in a story might have done something different but the outcome would remain the same, causal ratings for those actors decrease (Branscombe et al. 1996; McCloy and Byrne 2002; Spellman and Kincannon 2001).

Why should judging mutability before causality have this effect? Imagining alternatives that do not undo the outcome will create in the participants’ minds more ways in which the outcome would have occurred. The effect on Equation 2 is that such imagining will increase both the numerator and the denominator of the “probability before” [by increasing \( p(O_{before}) \) on the top and bottom by the same amount]. Thus, the value of that fraction will increase and the value of the “probability after” minus the “probability before” will decrease, resulting in less causality being attributed.

**Causality Before Mutability: When Should It Matter?**
The two examples above demonstrate that judging mutability first can affect subsequent causality judgments - but can judging causality first affect subsequent mutability judgments? Wells and Gavanski (1989) found that they did not. We, however, have created one situation in which they do and one in which they do not.

Causality judgments did affect subsequent mutability judgments when we (Spellman 2003) used a variation of N’gbala and Branscombe’s (1995) Experiment 1 story to investigate order effects. (Our story was based on the description in their article; it is closest to their immoral controllable condition.)

Participants read about Joe, a father, who was going to pick up his son, Jimmy, from school. On the way to his car, he stopped to talk to some people. Meanwhile, a neighbor drove by the school, waited with Jimmy for 15 minutes, and then offered to drive the boy home. On the way home, a drunk driver came out of nowhere and struck the car, and Jimmy was seriously injured.

Participants judged both mutability and causality. For the mutability judgments they were asked to complete the sentence: “The outcome of this event might have been different IF ONLY...” For the causality judgments they were provided with a list of people - Joe, Jimmy, Neighbor, Other Driver - and asked to rate from 0 (not at all) to 10 (very much) how much a cause of the outcome each of them was. Half of the participants judged mutability first; the other half judged causality first.

When the tasks were done in the same order as in N’gbala and Branscombe (1995), that is,
mutability listings before causality ratings, our results were similar to theirs: Joe was mutated most often and Other Driver was rated as most causal. However, reversing the order in which the tasks were done affected the mutability task (listing) but not the causal task (rating): participants again viewed the Other Driver as most causal but now he was most often mutated, too.

Why were the mutability judgments significantly affected by the causality judgments? When doing mutability listings first, participants are influenced by external availability from the story (primacy, focus, etc.). However, when participants do a causal rating task in which all of the names are provided by the experimenter, those names are now all highly available for mutating. The Other Driver was rated, on average, as much more causal than Joe. Thus, participants would have him available for mutating and they would notice that changing his actions would change the outcome in a big way.

In contrast to the mutability task, causality ratings did not change significantly as a result of order. The magnitudes of the causal ratings for Joe and the Other Driver were related to how the participants responded to the mutability question, but not to task order.

One might conclude, therefore, that order effects exist for both types of judgments. However, we believe that causality judgments do not affect subsequent mutability judgments when the effects of availability are removed. The literature has largely overlooked differences in measures. Mutability is almost exclusively measured by listings - looking at what a participant thinks up first. Causality is usually measured by ratings - looking at what a participant rates highest; importantly, sometimes the rated items are generated by the participant and sometimes they are
provided by the experimenter. We believe that tasks that involve "thinking up things" are more subject to availability than those that do not. (Consider the differences between recall and recognition tests.) Therefore, to unconfound listing/rating with mutability/causality we devised a new way to measure mutability - *mutability ratings on experimenter-provided alternatives* (Spellman 2001, 2003; note that Mandel 2003c uses mutability ratings on participant-generated alternatives).

Participants read a story that we have used in several other experiments:

A young woman was driving home from work. She had left early that day because it was a holiday weekend and traffic was very heavy. She was the first car to stop at a particular red light. Behind her was a long line of cars with a school bus at the end. As she was waiting for the light to turn green, she reached down to change the radio station. At that moment the light finally turned green, but she took an extra few seconds to find a song she liked. She then accelerated and the cars and bus accelerated behind her. Just as the school bus got into the intersection, a car driven by an upset man who had been fired that day came screaming through the red light from the other direction hitting the bus and injuring many children.

Typically, when asked to list something in the story that would change the outcome, participants are most likely to mutate something involving the woman: the time she left work, changing the radio station. When asked what caused the outcome, participants give their highest ratings to things involving the man: getting fired, running the red light (Spellman 2001, 2003).
In this particular experiment, however, we used ratings of experimenter-provided alternatives for both mutability and causality judgments. After reading the story, participants saw a list of twelve events from the story that had been generated by participants from previous experiments in response to either a mutability listing or causal listing question (e.g., the woman changing the radio station, the man running the red light, school being open that day). For each event, participants were asked to rate the extent to which they agreed or disagreed that the event was mutable and causal. All participants rated both mutability and causality with half the participants doing the tasks in each order.

On average, mutability agreement ratings were higher than causality agreement ratings for almost all events; in fact, the only items with positive causality agreement ratings (when those ratings were made first) were things involving the man (i.e., being fired, being in a bad mood, running the red light). More importantly, the order in which participants judged mutability and causality affected their ratings. The most interesting finding is that rating mutability first significantly increased (most) later causal ratings whereas rating causality first had only small effects on later mutability ratings.

Why did the causality judgments show task order effects? When judging mutability, one imagines alternatives to the event; such imagined alternatives should reduce the "probability before" estimates of the outcome occurring, and therefore, according to Equation 2, increase the later causal ratings of the events. Why did the mutability judgments not show task order effects? Usually (i.e., in cases not involving multiple sufficient causes), what people pick as causes are
already *but for* causes - things that, when changed, will undo the outcome. Thus, rating causality first (a) does not make any new events more available for mutation (as long as the mutation task also has experimenter-provided alternatives) and (b) does not provide information to any mutability-computation function.⁵

But why did mutability judgments show order effects in our replication of N'gbala and Branscombe (1995) but not in the "Bus" experiment? In the replication, mutability was measured by listings; thus, items provided by the experimenter for the causal rating task became more available for the mutability listing task. However, when all items are provided to the participants within the mutability task, and we measure ratings rather than listings (as in the "Bus" experiment), such "availability" has no meaning or effect.

**Phenomenon 3: Action and Inaction Effects in Regret Judgments**

Until now we have focused solely on mutability and causality judgments. We believe, however, that our equations can be used to explain the so-called “action” and “inaction” effects for feelings of regret. The “action effect” - the idea that people would feel more regret for actions than for inactions - was first described by Kahneman and Tversky (1982). In their classic experiment, participants read about Mr Paul and Mr George. Mr Paul owns Stock A. He thinks about switching to Stock B but decides against it. Later he discovers that he would have been better off by $1,200 if he had switched. Mr George owns Stock C. He thinks about switching to Stock D and decides to do so. Later he discovers that he would have been better off by $1,200 if he had *not* switched. When asked who would feel more regret for his actions, a majority of participants judged that Mr George would feel more pain.
Action or Inaction?

Subsequent attempts to explain or reduce the action effect considered various aspects of the scenarios and the judgments including: whether the alternative outcomes were known or unknown to the protagonist (i.e., do they know what would have happened on the road not taken), whether the outcome was good or bad, whether to consider the protagonist’s regret in the short or long run (see, e.g., Byrne and McEleney 2000; Gilovich and Medvec 1994; Landman 1987). Another breakthrough came when Zeelenberg and colleagues (Zeelenberg et al. 1998; Zeelenberg et al. 2000; Zeelenberg et al. 2002) suggested that regret for actions versus inactions should depend on whether it seems prudent (or not) to keep the status quo based on the outcomes of prior instances. For example, if a soccer team is having a winning season, a coach who changes his players and loses should feel more regret than a coach who keeps his players and loses (an action effect). However, if the team is having a losing season, a coach who keeps his players and loses should feel more regret than a coach who changes his players and loses (an inaction effect). Zeelenberg and colleagues argue that people experience regret when decisions appear abnormal relative to previous outcomes.

These results follow from our view of how regret is related to causal and counterfactual reasoning (although we don't believe that "abnormality" is the mechanism). We assume that regret arises as an after-the-fact emotion when, in retrospect, you believe that you had the power to choose a course of events that would have been more likely to lead to the desired outcome than the one you actually chose. Note several important factors: (a) regret requires a choice among actions, at least some of which would change the probability of the outcome (i.e., the
decision is causal); (b) regret results from a comparison of what one actually did to what one could have done (i.e., it involves comparing the "factual" to one or more counterfactuals); and (c) regret requires having chosen what seems to be (in hindsight) a less promising choice.

**Our Regret Data**

We have replicated Zeelenberg et al.’s (2002) action and inaction effects by manipulating the prior record of a baseball pitcher and measuring the coach’s post-decisional regret. Participants read about two baseball teams each with records of 10 wins and 10 losses. One coach decides to play his regular starting pitcher whose record is either 7-3 or 3-7; the other coach decides not to play his regular starting pitcher whose record is also either 7-3 or 3-7. Both teams lose.

Participants were asked to estimate (a) the probability of the team winning given the coach's actual decision and (b) the probability of winning imagining that the coach had made the counterfactual decision. Some participants estimated those probabilities in foresight (before learning that the team lost); others estimated those probabilities in hindsight (after learning that the team lost). All participants were also asked to judge how much regret each coach would feel about his decision (on a scale from 1 = no regret at all to 10 = completely regrets).

Table 1 shows some of the results from this experiment. Note that the percentages are expressed for the chances of winning, not losing (which was the actual outcome). Participants clearly believed that the decision to keep or change the starting pitcher would affect the team's probability of winning - as can be seen by comparing the foresight factual and counterfactual probability estimates.
The action effect can be seen in the top half of the table. When the pitcher had a winning record and the coach kept him, the probability of winning given the factual decision ($M = 58$) was estimated as higher than the probability of winning given the counterfactual decision ($M = 40$); thus, the coach would feel relatively little regret for his good decision ($M = 5.7$) despite the loss. When the pitcher had a winning record but the coach changed him, the probability of winning given the counterfactual was higher than given the factual; thus, the coach felt more regret ($M = 8.5$).

The inaction effect can be seen in the bottom half of the table. When the pitcher had a losing record but the coach kept him, the probability of winning given the counterfactual was again higher than given the factual; thus, the coach felt a lot of regret ($M = 7.1$). However, when the coach changed the pitcher with the losing record his decision was seen as better than the counterfactual decision, so his regret rating was small ($M = 6.3$).

We therefore see an action effect for winners (more regret when the pitcher is changed) and an inaction effect for losers (more regret when the pitcher is kept). Because some of these measures were collected between subjects, we are limited in which correlations we can run; further research will remedy that problem.

**Other Possible Extensions: The Hindsight Bias**

We believe that our proposal has consequences for understanding other related forms of
reasoning including the hindsight bias (see Nario and Branscombe 1995; Roese and Olson 1996). In hindsight bias experiments, participants are asked to estimate the left-hand probability of Equation 1 ("probability after"); that is, estimate that the actual outcome will occur given that the target event has happened. In the control condition, participants make that estimate without knowing what outcome will occur. In the experimental condition, participants make that estimate after learning which outcome (e.g., winning the game) has occurred - but they are supposed to make the estimate as if they had not yet learned the outcome. The hindsight bias refers to the finding that after learning the outcome, experimental participants cannot put themselves back into the “unknowing” state, and their "probability after" estimates for the actual outcome are higher than those of control participants.

The knowledge that an outcome has occurred fosters the belief that it was more likely to occur (compared to the control condition). Increasing the “probability after” for the actual way it happened in both the numerator and denominator will increase the value of the fraction in Equation 2. Counterfactual reasoning has often been proposed as one way to decrease the hindsight bias by attacking that change in belief in the likelihood of the actual outcome occurring. Sometimes debiasing is partially successful: participants who are asked to imagine ways in which the outcome might not have occurred often demonstrate a reduction of hindsight bias relative to those who did not imagine such alternatives (e.g., Fischhoff 1982); other times, however, attempts to debias by providing (more) alternatives seem to backfire (e.g., Roese and Olson 1996; Sanna et al. 2002).^6

We are just beginning to experimentally test the predictions that follow from our equations
regarding the hindsight bias. For example, our equations predict that some kinds of
counterfactuals should be more effective in reducing the hindsight bias than others (depending
on how they affect the various components in Equation 3). The equations also suggest that
estimates of the existence of unarticulated (i.e., not specifically generated) counterfactuals may
explain some of the “backfiring” effects.

CONCLUSIONS
We have argued that despite their differences, counterfactual and causal reasoning are related in
that: (a) they rely on similar underlying knowledge and information and (b) making one type of
judgment may provide information for (and thus affect) judgments of the other type. We have
described how for something to be called a cause, it need not provide an outcome-changing
counterfactual (i.e., it need not be a but for cause - although it usually will); yet we still believe
that counterfactual reasoning informs our causal reasoning. It does so in two ways. First, judging
mutability before causality may bring to mind, and make more available, the mutated item (e.g.,
action or event) for consideration in the later causality judgment. Second, according to our
equations, imagining ways in which an outcome could have been undone (or not) will affect our
beliefs about the likelihood of that outcome occurring, which, in turn, will affect our beliefs
about causation.

Making causality judgments first, however, should not affect mutability judgments unless the
experimenter provides items (external cues) for the participants to rate as to causality and then
asks the participants to generate mutations. Generating causes before mutations will not change
those latter judgments because causes, when undone, typically will change an outcome; thus, the
causes considered will be a subset of the mutations considered.

Our theory seems to capture several similarities and differences between counterfactual and causal reasoning. At present, we are trying to extend it to explain judgments of regret and disappointment and the hindsight bias. Indeed, because counterfactual and causal reasoning are so important for, and so ubiquitous in, thinking generally, we are optimistic that understanding them will give us insight into some of the other complex capabilities of human thought.
FOOTNOTES

1. This view is somewhat analogous to the view in the covariation literature (Cheng 1997; Cheng and Novick 1992) that a cause is something that increases the probability of an effect above its baseline probability (see Spellman 1996, 1997, for that argument).

2. It is not clear to us why that explanation is required by the results. There was a change made to the story and that change might have affected counterfactual and causal reasoning separately and directly without the former mediating the latter.

3. Our lab and several others (e.g., Rachel McCloy, personal communication) have had difficulty replicating the wine experiment 10 or more years later. Few participants mutate the boss's decision or agree that his decision caused the employee's death. These days, many participants note that people should know wine is a common ingredient in cooking; many state that the woman should have told the waiter about her allergy; some say that she should have been carrying drugs to combat her allergic reaction; others want to know why she would allow her boss to take her out to dinner, much less order for her.

4. When all participants are included, as shown in the text, the omnibus ANOVA is significant but only the 1-of-5 and 5-of-5 conditions are different from each other. This experiment was run within subjects and when participants who gave equal ratings to the three conditions are excluded, the means are all significantly different from each other: 1.4, 0.7, and -1.6. We believe that such participants were likely to think something other than the ordering decision, for
example, the restaurant, was the cause of the illness.

5. If anything, rating causality first should lower some mutability ratings, as people have now probably considered the true causal factors and rated those items as most mutable. See Spellman 2003.

6. Figuring out how to decrease the hindsight bias is of particular interest to lawyers whose clients have been sued for negligence (Kamin and Rachlinski 1995; Spellman and Kincannon 2001). A finding of negligence requires an after-the-fact determination of what should have been before-the-fact foreseeable.
AUTHOR NOTES

This research was supported by a grant from The US National Institute of Mental Health (NIMH) to the first author. Portions of this research were presented at the EAESP Small Group Meeting on Counterfactual Thinking, La Baume, Aix-en-Provence, France, in May 2001. We would like to thank Hayley Daglis for technical assistance.
FIGURE CAPTION

Figure 2.1. Illustration of the Relation between Counterfactual and Causal Reasoning
REFERENCES


Greene, E.J. and Darley, J.M. (1998) “Effects of necessary, sufficient, and indirect causation on


Table 2.1. Results from our Causality, Counterfactual Reasoning, and Regret Study.

<table>
<thead>
<tr>
<th>Record / Action</th>
<th>Probability of Winning Judgment</th>
<th>Regret (1-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factual</td>
<td>Counterfactual</td>
</tr>
<tr>
<td>Winning Record</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keep Starter</td>
<td>Foresight</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Hindsight</td>
<td>58</td>
</tr>
<tr>
<td>Change Starter</td>
<td>Foresight</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Hindsight</td>
<td>38</td>
</tr>
<tr>
<td>Losing Record</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keep Starter</td>
<td>Foresight</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Hindsight</td>
<td>34</td>
</tr>
<tr>
<td>Change Starter</td>
<td>Foresight</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Hindsight</td>
<td>67</td>
</tr>
</tbody>
</table>
External Cues
(e.g., description of events; framing of questions)

Internal Cues
(e.g., potential answers generated to earlier questions)

Availability of Alternatives
Pre-existing Knowledge
(especially causal knowledge)

$p(O_{after})$

$p(O_{before})$

Causal Judgments
$p(O_{after}) - p(O_{before})$

Counterfactual Judgments
(what gets mutated)