Gaze behavior during predictive and diagnostic reasoning

by

Duncan Yao Amegbletor

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 14, 2019

Keywords: causality, contingency, predictive inference, diagnostic inference, eye tracking

Approved by

Alejandro A. Lazarte, Chair, Associate Professor, Department of Psychology
Francisco Arcediano, Research Associate, Oakland University Department of Psychology
Martha Escobar, Associate Professor, Oakland University Department of Psychology
Daniel J. Svyantek, Professor, Department of Psychology
Jeffery S. Katz, Professor Department of Psychology
L. Allan Furr, Professor, Department of Sociology

Abstract

Many current theories of causal learning and reasoning offer causal structure as the answer to a long-standing question: what allows people to distinguish causal relationships from non-causal relationships? One component of causal structure is causal priority – the idea that causes precede effects – and that this implicit understanding of temporal order is integral to peoples' ability to reason about causal relationships. In experimental settings, causal order can be manipulated by presenting the cause and effect in two orders: a predictive order where cause antecedes the effect, and a diagnostic order where effect antecedes the cause. Manipulating the causal order often changes the properties of both learning and making inferences about causal relationships. The purpose of this dissertation is to investigate differences in reasoning in both predictive (cause-effect order) and diagnostic (effect-cause order) causal relationships. In two experiments, gaze behavior measured via eye tracking was used as a metric of processing during causal reasoning. Gaze behavior was selected specifically to build upon prior investigations of causal order effects that examined measures of response speed and difficulty. In the first experiment, human participants completed a causal judgement task on an eye tracker. The second experiment had participants complete a similar causal judgement task, however, some of the information on screen was obscured during judgement making. In Experiment 1, it was observed that participants engaged in more processing during causal diagnostic trials than during causal predictive trials. The results of Experiment 2 are mixed, but offer interesting possibilities for the future of causal order research.

Acknowledgments

Thank you to Drs. Martha Escobar and Francisco Arcediano for all of their time spent guiding me through my development as a scientist, and to Dr. Alejandro Lazarte and the rest of my committee for all of the thoughtful feedback they have provided. I would like to thank my colleagues Dr. Elizabeth Dunaway, Harrison Brown, Tyler Toledo, and Zebulon Bell, for their help in both preparing and testing the experiments, and running participants. I can offer nothing but immense gratitude to my parents, Duncan Amegbletor Sr., and Lauretta Amegbletor for cultivating in me a lifelong commitment to learning and critical thought. I'd like to thank Nico Amegbletor, Dr. Andrew Shen, Mostafa Foda, and Dr. Derek Pope for their invaluable friendship, stimulating conversations, and refuge from the trials of my time in graduate school. Finally, I am deeply indebted to Sharay Setti for her unconditional love and unyielding support without which I would not have completed this grand undertaking.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Learning to make inferences about the relationships between events in the environment is one of the building blocks of human cognition. Causal reasoning allows humans to predict negative consequences while seeking out desired outcomes, to exert control over events in the environment, and to provide the basis for the scientific and philosophical knowledge we have used to understand and explore the universe. Assessing the world through a causal lens is a consistent part of the human experience. In antiquity, humans explained natural phenomena (e.g., sunrises and tides) as being caused by the actions of spirits and deities. As humanity began to intervene on the environment through technology and agriculture, our perception of what could be a causal agent shifted to the material realm. The shift in our perception of causes led scholars in ancient Africa, Asia, and Europe to develop ideas of how quantifiable causes and effects interact, and to theorize about systems of causality (Gopnik, 2009; Pearl, 2000). Aristotle described four causes – efficient, material, formal, and final causes – as the four levels of explanation for the natural world (Barnes, 1991). Among the four causes, efficient causes are what we typically think of as causality; efficient causes act to trigger or prevent changes in objects and states. Aristotle alluded to the existence of two types of efficient causes: generative causes (causes producing an effect) and preventive causes (causes preventing an effect). For Aristotle, a sculptor acts as an agent generating change by chiseling away a statue from a slab of marble. Likewise, a woodworker can be an agent preventing change by applying a varnish to wood and protecting the material below from water damage. Causal connections can be simple (one cause connected to one effect), or complex (e.g., causes that produce effects that in turn produce other effects, or multiple causes contributing to the same effect). Regardless of the

type of connection, people tend to produce causal inferences about two events when they are correlated in space and time, (i.e., when they occur in close spatial or temporal proximity).

When learners consider an event as an effect of a cause, either the cause on which the learners are focusing (the target cause) or another alternative cause (a background cause) that is not under current focus is the correct one. For example, if when playing billiards, a ball moves without being struck by another ball, it is natural to assume that some cause other than collision (e.g., a gust of wind or a tilt of the pool table) caused the movement. If the effect occurs in the absence of the target cause, the contingency between the target cause and the effect weakens, while the contingency between the effect and background causes strengthens. Similarly, if the target cause occurs without producing the effect (i.e. a ball struck another one without moving it) the association between the target cause and the effect weakens. This shift in association happens even when the background cause is not readily apparent. Thus, in order to maintain a strong connection between a target cause and an effect, the two events must frequently co-occur together, and they should rarely occur apart. Events which occur physically close together and/or in close time proximity tend to be more easily associated than events which are distal in space and time. This idea of spatiotemporal contiguity is one of Aristotle's laws of association – rules which govern how we acquire knowledge (Barnes, 1991). Illness that occurs days after eating contaminated food is harder to attribute to the proper source than illness that occurs a few hours after eating. Likewise, if a billiard ball moves several seconds after being struck, an observer will look to other potential causes rather than the collision of cue and ball.

Causal priority (also referred to as causal order or temporal priority) refers to the fact that the cause temporally precedes the effect. This directionality of the causal relationship is important for intervening on the environment to produce or prevent effects. However, the order in which we observe the cause and effect may affect our behavior. Take, for instance, a person who has intense sneezing fits in the presence of cats and goes to visit a friend with cats. The

knowledge that the cause precedes (and triggers) the effect allows the person to intervene on the cause rather than the effect. Thus, knowing that cat dander is the cause of the sneezing will lead our hypothetical allergy sufferer to prevent the action of the cause (e.g., take an antihistamine or minimize contact with the cats during the visit), rather than trying to act on the effect. However, people can also observe an effect without first observing or being aware of a cause; after observing an effect, people tend to generate potential causes for that effect. If our hypothetical allergy sufferer begins sneezing uncontrollably after entering a new acquaintance's house, they could make the inference that there are cats in the house before observing any feline presence. Making inferences about potential causes when first observing an effect is referred to as diagnostic inference or diagnosis. In diagnostic inference, the effect is present because the cause has already occurred, or is still occurring, at the time of inference making. On the other hand, making inferences about potential effects (that have yet to occur) after observing a cause, is referred to as predictive inference or prediction.

Hume, Kant, and the problem of induction

In *A Treatise of Human Nature* (Hume, 2003), the Scottish philosopher David Hume describes the properties of efficient causes (henceforth referred to as causes), and the difficulties in inferring causal relationships that cognitive systems must overcome. At the center of Hume's empiricism is the copy principle – the concept that all elements of human cognition come from experience. Everything we think about, know, process, etc. comes from our observations of the world around us. Causality is no different; we require experience in order to identify which events are potential causes, and what events are likely to follow those causes as effects. The second premise of the copy principle is that we can recombine experiences into new ideas. With causality, this allows us to infer novel causal relationships based on resemblance to past experience. For instance, a person who has never seen a game of pool can still predict the trajectories and speed of the balls based on their prior observations of

moving objects; likewise, a child can decide not to touch a hot iron after prior experience with being burned.

We make these causal conclusions based on the observation of co-occurring events. However, this correlation between events does not by itself necessarily establish that the relationship between said events is causal. Frequent co-occurrence of two contiguous events can be misleading – the two events could share a predictive yet non-causal relationship, or co-occurrence could be due to the existence of other unseen causes impacting one or both of the events. The inherent difficulty in elucidating causality from observed events is at the core of an integral component of Hume's system of causal inference: the problem of induction. Hume maintains that causality as a force is not directly observable; what we do observe is repeated conjunction between certain events. From this information about the conjunction of events we make the assumption that Event A causes Event B. Hume, however, objects to the idea that there is some sort of causal necessity or causal power in these observed relationships. Instead he argues that causality is a mental construct, an idea that arises from observations of events that frequently co-occur.

In response to Hume's criticism of causal inference, Kant (1848) not only attempted to reify our sense of causality, but portrayed it as an indelible aspect of the way that humans perceive the world. Kant agreed with Hume that causal powers are not directly observable, and that the problem of induction provides a convincing skeptical challenge to old ideas of causality. Rather than concluding (as Hume did) that causality is a "habit of the mind", causality to Kant was an *a priori* concept: causality is a perceptual lens that shapes the way we observe and understand the world. Other *a priori* concepts (e.g., our perception of time and space) work in the same manner in Kant's epistemology. Causality is required to make sense of the world, to differentiate between "sequences of states" and "sequences of perceptions". As an example, Kant explains the difference between observing a stationary object (a house) and an object in

motion acted on by various forces (a ship). We can perceive and re-examine the elements of the house (the roof, the windows, the lawn, etc.) in any order. The movement of the ship, however, is linear, we cannot go back and perceive the ship in a previous, upriver state. Rather, the state of the ship at the current moment is an effect of the previous states of the ship. All observations, whether they are causal or not, enter our senses as a "sequence of perceptions". Our *a priori* sense of causality is required to make successive changes in state meaningful, and to make causal inferences valid.

Together, Hume and Kant provide the intellectual foundation for much of the empirical investigation of causal learning and inference from the 20th century to the present. However, the debate over which system of causality provides a better fit to human causal inference is not yet resolved. Hume's view of causality holds that the way we understand causality is a mental construct. Observation of constant conjunction is all we have and the necessity of causality is what is constructed by the mind to explain our observations. Hume's perspective is one of strict empiricism – sensory observation is the only certainty. Kant's view proposes a different answer to the problem of induction. Rather than causality being truly unknowable regardless of how many experiences one has, Kant believed that the idea of causal necessity (causes must produce effects) is an *a priori* assumption of the mind. Two modern day perspectives of causal inference are the successors to the ideas of these two philosophers.

Modeling Causality. Contemporary cognitive models of causal learning, such as Causal Model Theory (Waldmann & Holyoak, 1992) and various Bayesian learning models (e.g., De Houwer, 2009; Gopnik et al., 2004), assume the Kantian variant of causality and thus place importance on *a priori* causal structures shaping how we observe and identify causal relationships. Causal structure refers to the idea that humans have an inherent knowledge of how causal relationships are structured.

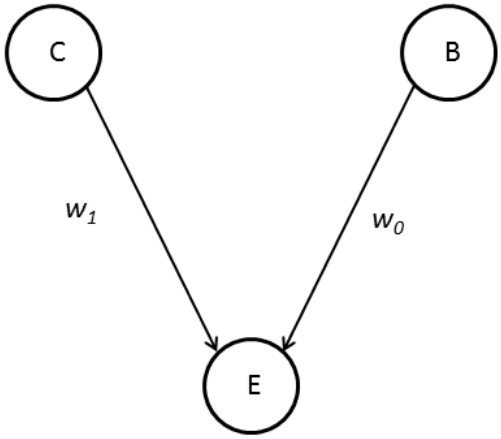

Figure 1. An example of a causal map. Here both the target (C) and alternative (B) causes can produce the outcome (E). W_1 and W_0 are weights, which represent the strength of the causal relationship between each individual cause and the effect.

These cues of causal structure are features (e.g., contingency/covariation, causal order, resemblance to prior knowledge of causal relationships, ability to intervene on the relationship) that indicate the presence of causal relationships (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). For example, developing a rash after eating shellfish meets several of these criteria for causality: (1) The cause and effect occur in temporal proximity, with the rash (the effect) occurring after eating shellfish (the cause); (2) The relationship in question resembles other occurrences of foods causing bodily reactions, and (3) taking an antihistamine prior to eating shellfish (intervention) can prevent the rash. *A priori* models of causality propose that causes

and effects established to have a structured causal relationship are encoded into mental representations of the observed relationship. These mental representations are referred to as causal models or causal maps. Figure 1 depicts an abstraction of causal map; here, both the target cause (e.g., eating shellfish) and a background cause (e.g., eating an avocado) can produce the effect. In Figure 2, the target is a non-causal event and the effect is produced by the background cause.

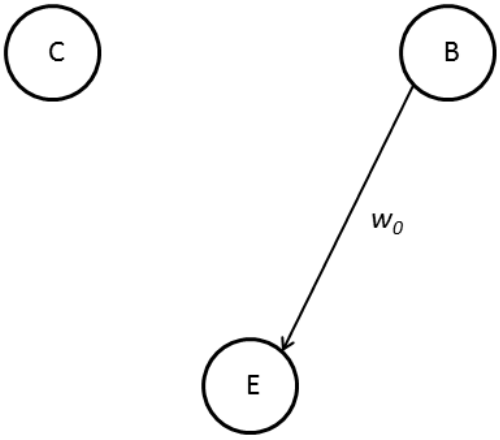

Figure 2. An example of a causal map. Here only the alternative cause (B) produces the outcome (E), the target event is non-causal. W_0 is a weights which represents the strength of the causal relationship between cause B and the effect. Event C is not a cause of the target event, and thus has no weight.

In *a priori* models, causes and effects have distinct roles – every effect is preceded and produced by some cause, and this temporal order is encoded as part of the causal map. Because the direction in which causality flows is encoded into the mental representation, asymmetry of causal order is ingrained into *a priori* models. When learning a relationship from effect-cause information, the input is stored in a cause-effect ordered mental representation.

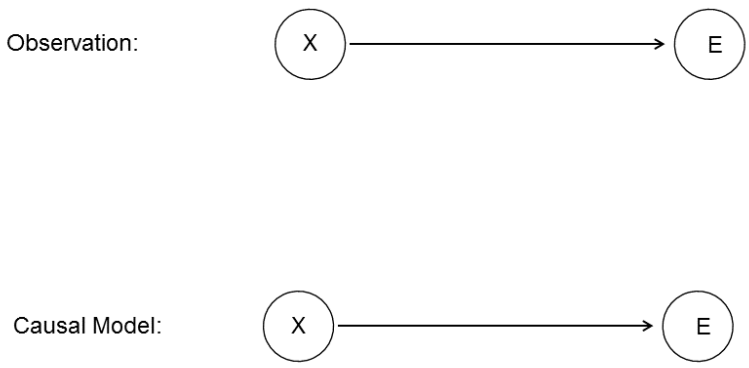

Figure 3. Predictive inference in Causal Model Theory. The target cause (X) precedes and produces the effect (E). Here the observed events and causal model are congruent.

Even though causal models are unidirectional, people can make both predictive and diagnostic judgements with this knowledge. To make inferences from diagnostic information, the information needs to be "rotated" in order to match the previously stored cause-effect mental representation. Thus, reasoning from cause-to-effect differs from reasoning from effect-to-

cause. See Figures 3 and 4 for examples of predictive and diagnostic causal reasoning.

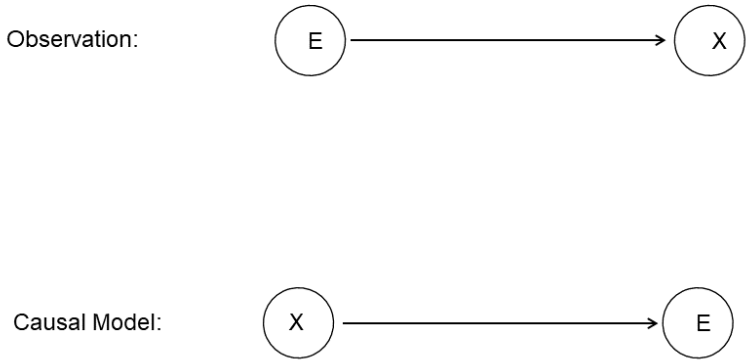

Figure 4. Diagnostic inference in Causal Model Theory. The observed effect (E) precedes the target cause (X). However, the causal model reflects that X is the cause, even though it is a subsequent event.

Because of the discrepancy between mental representation and observation in diagnostic reasoning in *a priori* models, diagnostic inference should take longer to achieve than predictive inference (e.g., Fenker, Waldmann, & Holyoak, 2005), employ information about other causes differently than in predictive inference (e.g., Fernbach, Darlow, & Sloman, 2010), and be more cognitively effortful than predictive inference making (e.g., Bright & Feeney, 2014a).

In contrast to *a priori* models of causality, reductionist models like the contingency rule delta-P (ΔP; Allan, 1980; Jenkins & Ward, 1965) take the Humean assumption and thus reduce the construct of causal inference to not-explicitly-causal associations built out of the contingency or correlation between events. Accordingly, a learner's behavior and judgements of the relationship between a cause and effect should reflect the value of the contingency between

that cause and effect (this same assumption applies to non-causal relationships). Given a potential relationship between two events, there are a finite number of configurations in which they can co-occur. A 2x2 matrix (see Table 1) is often used to describe the four possible ways a cause and an effect can covary, and records the frequencies of each type of joint occurrence.

Table 1

The ΔP 2x2 Contingency Matrix

	Effect	
Cause	E	~E
C	a	b
~C	c	d

As an example: the table could be used to determine the extent to which eating shellfish (cause) produces facial swelling (effect). In this example, cell a would represent occasions when one ate shellfish and developed facial swelling, cell b represents occasions where one ate shellfish, but had no facial swelling, cell c represents occasions when one had the allergic reaction without eating shellfish, and cell d represents occasions when one did not eat shellfish, and had no reaction. This 2x2 table serves as a visual summary of a learner's experience with a potential causal relationship. Using the ΔP rule, the strength of the causal relationship would be determined by subtracting two conditional probabilities: the proportion of disconfirmatory events (the probability of occurrence of the effect in the absence of the cause, P(E|~C) or (c/c+d)) from the proportion of events that confirm the relationship (the probability of the occurrence of the effect given the cause, P(E|C) or (a/a+b)). That is, the contingency between events (ΔP) would be represented by P(E|C) - P(E|~C) or (a/a+b) - (c/c+d).

The point of view one takes about explaining causal inference affects how predictive and diagnostic inference are treated. In the pure contingency models, where there is no explicit use of causal structure in a model, and causal and non-causal relationships are encoded in the same cue-outcome framework: causal order manipulations are not relevant as the associations are not directional. On the other hand, in *a priori* models, causal structure determines how relationships are inferred, causal order is of high importance and manipulations will significantly affect causal induction. The following section will analyze the evidence surrounding the idea that causal order manipulation has an effect on human causal learning and inference.

Causal order effects

For centuries, the nature and mechanisms of causality were topics of debate among philosophers (e.g. Aristotle, Hume, Kant), but over time, the investigation of causality became a scientific endeavor. Within the study of causality, the differences between predictive and diagnostic inference are of recurring interest. The consensus appears to point towards some degree of asymmetry in predictive and diagnostic inference making. However, questions regarding the conditions under which we see differences, and the impact this dissociation ultimately has on causal learning and reasoning remain open to discussion

Consideration of alternative causes. Normative expectations of causality (e.g., all effects require a cause, causes precede effects) give a basis for differences in the way information is used in causal reasoning. Experiments investigating causal reasoning typically present a scenario in which events share some pattern of relation, then request participants to make a judgement about how strongly they think one event is caused by another, or how confident they are that a causal relationship exists between events. Participants' judgements are typically quantified as rating on a numerical scale (e.g., 1-10, or 1-100) with larger numbers

representing more certainty about the causal relationship, or as a dichotomous response as to whether a causal link exists or not.

In predictive reasoning, alternative causes (causes other than the target cause) do not compete with the target cause or each other – here the question commonly posed is whether the effect will occur given these causes. Both the target and alternative causes increase confidence about the occurrence of the effect. For instance, being in the presence of both a cat, and fresh roses (assuming that the cat allergy sufferer also has a pollen allergy) only magnifies the chances of a sneezing fit occurring. In contrast, alternative causes compete with each other in diagnostic reasoning – the question being whether the observed effect (the sneezing fit) was produced by a specific cause (the cat versus the roses). Multiple potential causes may reduce the likelihood that the specific target cause produced the reaction (Fernbach et al., 2010; Pearl, 1988; Waldmann & Holyoak, 1992). Imagine our allergy sufferer visits a botanical garden (but forgets to take any medication); here the sheer number of different pollens makes it unclear which specific plant is causing the reaction. Looking at judgements normatively, we would expect that when alternative causes are present people inflate their predictive judgements (consider the effect much more likely to occur), and deflate their diagnostic judgments (consider the target cause much less likely to be the true cause of the effect). Classic studies (e.g., Tversky & Kahneman, 1980) report that people tend to judge the probability of predictive ordered statements occurring as higher than that of diagnostic ordered statements, even when all other factors (besides the order of events in the statement) are controlled. Tversky and Kahneman (1980) had participants select which of two causal descriptions they thought was more probable (or state that the two options were equally probable). As an example, the authors asked participants which of the following was more probable: (a) That a girl has blue eyes if her mother has blue eyes, or (b) That the mother has blue eyes, if her daughter has blue eyes. More participants selected the cause-effect ordered option (a) than the effect-cause ordered

option (b). Participants' perception of the strength of a relationship differed even in predictive and diagnostic presentations of the same scenario. However, more recent research suggests that there is significant departure from normativity (i.e., predictive judgements are not always higher) when it comes to the effects of alternative causes. The presence of alternative causes (in the diagnostic direction) will often lower the perceived causal strength (Rehder & Waldmann, 2017; Waldmann, Hagmayer, & Blaisdell, 2006).

Fernbach, Darlow, and Sloman (2010) compared predictive and diagnostic judgements of full conditional statements (where alternative causes are present) and no-alternative conditional statements (where there are no valid alternative causes). Participants were given questionnaires containing predictive or diagnostic framed causal scenarios. Each scenario was written in two forms: full conditional (where unstated alternative causes could have also produced the effect) and no alternative (where all alternative causes other than the target cause were eliminated). The statement "Ms. Y is a 32-year-old female who has been diagnosed with depression." is an example of a full conditional. Adding the sentence "A complete diagnostic workup reveals that she has not been diagnosed with any other medical or psychiatric disorder that would cause lethargy." converts the first statement into a no-alternative scenario. Participants rated the likelihood of the effect (lethargy) given the target cause (depression) in predictive framing or the likelihood of target cause (depression) producing the stated effect (lethargy) in diagnostic framing on a scale of 1-10. Given the expectations of how alternative causes affect reasoning, alternatives should increase the likelihood of the effect occurring in predictive reasoning, and participants should give higher ratings to full conditionals. In diagnostic reasoning, full conditionals should be rated lower than no-alternatives. Because participants are judging whether effect Y occurred because of cause X, more potential causes make the role of X less clear, and thus the ratings for the target cause should be lower in the diagnostic full conditional. Fernbach et al., (2010) found that, as expected, no-alternative

diagnostic judgements were rated higher than full conditional diagnostic judgements. However, for predictive judgements, there was no difference between full conditional and no-alternative statements. The authors took this to indicate that people neglect the role of alternative causes in predictive scenarios, causing them to underestimate the likelihood of the effect when alternative causes are present – counter to normative expectations. No such bias seems to occur in diagnostic reasoning. In a separate paper, the authors found that a Bayesian network model correctly tracked participant performance for diagnostic judgements, but failed to capture participants' predictive judgements, due to non-normative causal neglect (Fernbach, Darlow, & Sloman, 2011). Fernbach et al., (2010, 2011) provide convincing evidence that people use causal structure in a non-normative way. However, just as they affect responses during causal learning, task instructions and the phrasing of the request for a causal judgement have a large impact on whether causal reasoning is found to be non-normative.

Fernbach and Rehder (2013) presented participants with a scenario in which target features caused other features in the members of a category. For example, in the fictional category "myastars", the high density of myastars causes them to have a large number of surrounding planets. In Experiment 1, alternative causes of the effect were implied to exist; participants were informed that "other features of myastars" were capable of causing the effect (large number of planets). In Experiment 2, implicit alternative causes were compared to alternative causes that were explicitly stated – one of the features of the category (e.g., the high temperature of myastars) was stated to be an alternative cause of the effect. The probability of each target cause (e.g., high density) and alternative cause (e.g., high temperature) was manipulated to make the category feature either a weak or strong cause of the effect (e.g., large number of planets). Participants were asked to make predictive and diagnostic confidence judgements on the occurrence of the features. Fernbach and Rehder (2013) reported that predictive ratings of causal strength in the categories were affected by the strength of the target

cause, but insensitive to the strength of alternative causes (even when alternative causes were explicitly given). In contrast, diagnostic inferences were sensitive to both the strength of target and alternative causes. This finding further supports the neglect of alternatives in predictive inference. However, when the alternative cause was stated to be present/true in the question prompting the participant to make a causal inference, alternative strength was correctly considered in predictive judgements. Explicit, unambiguous information seems to make reasoners less likely to take shortcuts that results in errors (e.g. neglecting alternative causes). In summary, the slice of literature examining people's use of information regarding alternative causes strongly supports an asymmetry between predictive and diagnostic inference making. Alternative causes are important to making diagnostic inferences, and people treat alternatives as irrelevant in predictive inference despite the normative expectation of inflation in predictive ratings.

Development of causal structure. Additional evidence for an asymmetry of causal order can be found in studies investigating children's ability to apply causal knowledge. Bright and Feeney (2014a) investigated whether children (5, 8, and 12-year-olds) display causal asymmetry, which would arise from using the same underlying causal structure used by adults. On each trial, participants were shown a base image (e.g. a picture of carrot) and two target images. One of the targets was causally related to the base (e.g., horse eats carrot) and one target was taxonomically related to the base (e.g., carrots and onions are both vegetables). Participants were told that the base object held a specific property (either a special type of cell or a specific disease) and they had to choose which of the two targets was more likely to possess that property. Participants of all ages tended to select causal targets over taxonomic targets on predictive trials. On diagnostic trials, this bias was less pronounced. This was discussed as indicative that predictive reasoning lends itself to causal inferences more easily than diagnostic reasoning, and the authors concluded that causal reasoning ability and causal

asymmetry appear early in development (around the age of 5). Fernbach, Macris, & Sobel (2012) examined causal inference making in 3 and 4-year old children. The children were shown that placing certain colored blocks on a “machine” caused it to play music. For predictive inference, the children were asked which block would make the machine work. For diagnostic inference, the machine was activated behind a screen, then the children were asked which block made the machine work. The authors found that both 3 and 4-year-olds have the ability to make diagnostic inferences and maintain belief that other alternative causes could have produced the effect (roughly 90% of trials were error free). However, only 4-year-olds were able to solve problems in which an uncertain cause served as the best explanation of the effect (about 44% of trials were error free; chance was at 16.7%). This line of research provides evidence that important elements of our causal reasoning ability may not be fully developed in early childhood. The slower development of diagnostic reasoning – as compared to predictive reasoning – provides interesting parallels to adult causal reasoning, where diagnostic reasoning seems to be more resource intensive than predictive reasoning.

Causal order in learning. A good deal of what is known about the differences between predictive and diagnostic inference comes from studies of stimulus competition in causal learning. Stimulus competition refers to a reduction in responding to a target cue as a result of a second cue serving as a better predictor of their shared outcome. The response of interest could be the ratings of the association between a potential cause and an effect in human causal learning preparations. Ivan Pavlov (1927) is credited with the earliest report of stimulus competition in what he termed “overshadowing”. Overshadowing usually involves pairing a more salient cue, A, in compound with a less salient cue, X (AX-Outcome). The less salient cue (X) typically elicits less responding, consistent with the A-Outcome association – e.g., someone who gets sick from eating Szechuan shrimp and rice, will likely attribute the illness to the more salient spicy food rather than the less salient rice. “Blocking” (Kamin, 1968, 1969) is also

commonly used in investigating causality. In blocking, a cue, A, is trained with an outcome, O (A-Outcome). Subsequently, A is trained in compound with a second cue, X, and followed by the same outcome (AX-Outcome). Blocking training typically results in diminished responses to the added cue, X, as compared to a cue with a similar training history – e.g., cue Y is paired in compound with cue B, then cue Y is tested along with cue X as a comparison. Cue A is typically not used as a comparison for blocking because it should elicit higher ratings than cue X due to its training history as a sufficient predictor of the outcome when presented alone (individual trials in Phase 1, then compound trials in Phase 2). Cues X and Y have identical training histories, the sole difference between them is that X has been paired with a reliable cause of the effect, while Y has been paired with an equally novel cue. Blocking occurs because the added cue (X) is redundant in predicting the outcome – e.g., someone who has learned that avocados give them heartburn, will likely attribute the heartburn to the green fruit after eating an avocado and tomato salad. Other types of stimulus competition exist (e.g., Escobar, Matute, & Miller, 2001; Matute & Pineño, 1998; Wagner, Logan, & Haberlandt, 1968), but are not relevant to the current discussion as they are not commonly used in investigating causal order effects.

Waldmann and Holyoak (1992) presented participants with either a predictive scenario (in which certain facial features elicited emotional responses in fictional observers) or a diagnostic scenario (in which specific symptoms indicated the presence of a virus). Each trial in the task was framed as the experience of one fictional person or patient. Participants observed 24 target trials in which the trait “pale skin” was established as an effect or a cause. The target trials were intermixed with 24 non-target trials. Afterwards, participants observed another 24 target trials in which the trait “pale skin” was paired with the trait “underweight”, again intermixed with non-target trials. At test, Waldmann and Holyoak observed a blocking effect in the predictive task (higher ratings to “pale skin” than to “underweight”), however competition did not occur in the diagnostic task. The authors proposed that stimulus competition in causal scenarios

is driven by abstract rules of causality rather than differences in the associative strength of the cues. The competition seen in the predictive direction was due to ambiguity about the added cue's role as a cause – it was only presented in compound with the previously trained cue so its individual causal power was unknown. In the diagnostic task, multiple effects of a common cause do not compete. A cause produces multiple effects independently – the presence of Effect A has no bearing on whether the cause also produced Effect X. Blocking in diagnosis can be caused instead by competing causal models – when a participant is asked about the occurrence of an effect, any valid (non-target) cause could lead participants to inflate ratings of that effect. Accordingly, eliminating competing causal models wipes out blocking in diagnosis (Experiment 3). There were however conflicting reports on stimulus competition in diagnostic inference that contradict Waldmann and Holyoak's conclusions. Chapman (1991) observed blocking in a scenario where participants used symptoms (effects) to diagnose fictional patients with a disease (cause). Kruchke and Blair (2000) observed blocking in a diagnostic causal reasoning task (symptoms as effects of fictional diseases) even when the learning phases of blocking were reversed (AX-Outcome, then A-Outcome). Shanks and Lopez (1996) replicated Waldmann and Holyoak (1992) while addressing methodological issues (e.g., low power, and the use of an inappropriate test for blocking) and consistently observed blocking in both predictive and diagnostic scenarios.

However, not all evidence in the field contradicted the findings of Waldmann and Holyoak (1992). Using a task in which lights on the front of a box potentially control a light on the back of a box (predictive) or where the lights on the front of the box are effects of the light on the back (diagnostic), Waldmann (2000) observed a lack of blocking in diagnosis, while attempting to address the methodological criticisms of Waldmann and Holyoak (1992). Waldmann (2001) found a significant difference between predictive (substances in blood are the causes of a disease) and diagnostic (substances in blood are effects of a virus) overshadowing;

however diagnostic inference making did not eliminate overshadowing. Kloos and Sloutsky (2013) presented children and adults with a scenario in which animals who ate magical fruit transformed into objects. In a follow-up experiment, the task was changed to a scenario where they were shown animal footprints and had to decide which animal had produced the prints (diagnostic) or whether the presence of the prints would scare away prey (predictive). Across the different tasks it was observed that children ignore causal structure in blocking, while adults use causal structure in the expected manner – accordingly, blocking in diagnostic causal inference occurred in children, but not adults.

Matute, Arcediano, and Miller (1996) attempted to resolve the controversy over claims of diagnostic stimulus competition. Rather than causal order being a strong global determinant of whether competition is observed, factors such as the causal order of the training, the casual order inherent to the question asked of the participant at testing, and whether the competing cues were causes or effects (independent of causal order) play a large role in participant responses. Information was presented to participants as a table in predictive (Experiment 1; Column 1: patient, Column 2: medicines, Column 3: allergic reaction) or diagnostic order (Experiment 2; Column 1: patient, Column 2: allergic reaction, Column 3: medicines). The participants were asked four questions (on a scale of 1 - 8) for each target cue – two to assess causality, two to assess contiguity between the cause and the effect. Competition among causes was not affected by the causal order of training or testing, while competition among effects was strongly dependent on the type of question posed to participants. Questions about causality did not elicit competition between effects (regardless of the causal order of the causal question). Competition between effects occurred in Experiment 3 when participants were asked "indicator questions", for instance – "Is effect A indicative that cause X has occurred?" Matute et al. hypothesized that the indicator and causality questions are asking participants to give different probability judgements. Asking about $P(C|E)$ compared to $P(C|\sim E)$ as in indicator

questions, promotes competition among effects (but does not eliminate competition among causes). The authors observed that prior studies observing competition among effects used questions in this vein. Arcediano, Matute, Escobar, and Miller (2005) investigated training and testing factors (type of cue, causal order at training, causal order at test) with the addition of whether the competition (overshadowing) is between antecedent or subsequent events. Participants saw 48 trials of information in both predictive and diagnostic order about fictional patients, the food(s) they ate, and what adverse reactions the patient developed. Using indicator questions, competition occurred in both predictive and diagnostic conditions regardless of the type of manipulation. López, Cobos, and Caño (2005) also hypothesized that the question posed to participants affected whether or not competition was observed. Specifically, when participants are asked about the causal roles of the cues and outcomes, stimulus competition is not observed in diagnostic inference making (e.g., Waldmann, 2001). However, when participants are asked to give a judgement of the "predictive or diagnostic value" of a cue, stimulus competition occurs in both predictive and diagnostic inference (Cobos, López, Caño, Almaraz, & Shanks, 2002). The experiment used a modified version of the abstract causal learning task created by Waldmann (2000) in which lights on the front of a box caused lights on the back of the box to light up (predictive), or lights on the front of the box were effects of lights on the back of the box (diagnostic). López et al., (2005) found empirical support for their predictive/diagnostic value hypothesis and were able to influence (or eliminate) stimulus competition in diagnostic inference by manipulating the task instructions.

The causal stimulus competition literature provides strong evidence that prediction and diagnosis are not symmetrical. Humans do use causal information flexibly, however, early research overstated the degree of causal asymmetry. Diagnostic reasoning does not necessarily eliminate blocking and other stimulus competition effects. Rather, factors such as the causal order of training and test, and the phrasing of the question posed to participants

control inference making. Further, diagnostic reasoning seems to be more sensitive to the specific probability being assessed than predictive reasoning. Questions relating to the mechanism that humans use to learn causal relationships, and how instructions and framing interact with this system remain open ended.

Making inferences is cognitively demanding. Evidence suggests that the process of making causal inferences – generating a mental representation of the relationship, considering the role of target and alternative causes – is cognitively effortful (De Houwer, 2009, 2014; Mitchell, De Houwer, & Lovibond, 2009). For instance, taxing a learner's cognitive resources with an irrelevant secondary task inhibits their ability to learn causal relationships (De Houwer & Beckers, 2003; Waldmann & Walker, 2005). Similarly, giving a learner more relationships to learn increases response time (RT), and providing very short response intervals can prevent inference making (Sternberg & McClelland, 2012). Bright and Feeney (Experiment 2; 2014b) examined this premise of cognitive effort in predictive and diagnostic causal judgements. The authors employed a distractor task during inference making in which participants needed to remember dot patterns of varying complexity (heavy load vs light load condition). The experiment used the same category task from Bright and Feeney (2014a). Predictive trials were given a higher causal rating than diagnostic trials in the light load condition, however this causal asymmetry was no longer significant in the heavy load condition. Fenker, Waldmann and Holyoak (2005) examined the ability of participants to retrieve previously learned causal relationships. Participants were asked to make a judgement about whether a pair of words was causally related (e.g., virus and epidemic) or associated, but not causally linked (e.g., graduation and gown). Both causally linked and associated word pairs could be presented in cause-effect (predictive) or effect-cause (diagnostic) order. The authors hypothesized that making diagnostic judgements would take longer (in terms of reaction time or RT) than predictive judgements. Both the order of observed events and the mental representation of the

relationship are congruent for predictive judgements; as such, retrieving and responding to a predictive relationship should be a relatively fast process. In contrast, for a diagnostic judgement the temporal order and mental representation are reversed. The incongruence of the word pair and the mental representation of the relationship must be resolved before a response can be made. The authors observed faster RT for predictive judgements than for diagnostic judgements. This effect was dependent on the word pairs being causally-related: the order of word presentation did not affect word pairs that were not causally related.

The literature reviewed thus far provides evidence that predictive and diagnostic causal inferences are asymmetric in the use of alternative causes, the way they develop in early childhood, and the speed and degree of cognitive effort induced by causal order. Further, the acquisition of predictive and diagnostic associations may also differ. The evidence suggests that the human reasoner does possess *a priori* knowledge of how causes and effects should interact. However, despite the wealth of behavioral evidence surrounding the *a priori* hypothesis, there is far less evidence directly supporting the assumption that reasoners use causal maps, and more specifically, the additional processing required to re-map diagnostic causal information. To further the science of causality, it is important to shed light on the covert processes of building and accessing causal maps. Eye tracking is a methodology that has proven fruitful in uncovering parameters of mental processes (e.g., attention). The subsequent section will briefly discuss eye tracking and its use in measuring cognitive processes such as human learning and reasoning, and segue into Chapter Two which presents two experiments designed to use gaze behavior (as measured by eye tracking methodology) to test the effect of causal order manipulation on causal reasoning and observe evidence of participants re-mapping diagnostic information.

Eye tracking

The experiments reported in Chapter II employ eye tracking technology. Eye tracking provides a means of directly measuring where a viewer's gaze lies. An eye tracking apparatus monitors stimuli and locations at which a person is currently looking. To understand the idea behind the use of eye tracking, it is important to understand exactly what eye trackers measure. In short, the first question to be addressed involves the form and nature of eye movements.

Eye movements. Saccades are quick and relatively short movements that bring target stimuli into focus, so that they may be viewed clearly and accurately. Typically, these shifts in gaze location cover 15-20 degrees of visual angle. Saccades allow us to scan the visual field for stimuli, and shift the fovea between different targets. Both eyes move simultaneously during a saccade, and the movement is ballistic – meaning that once the movement is initiated, one cannot stop it (Castelhano & Rayner, 2008). Visual perception is suppressed during a saccade; the velocity of the movement is too quick for accurate perception of stimuli (Dodge, 1900; Matin, 1974). Because vision does not occur during saccades, visual perception occurs during eye fixations.

Fixation is defined as the moment when the fovea (the region of the retina responsible for acute vision) is directed at a specific target allowing visual perception to occur. Despite the implication of the name "fixation", there are a number of movements that occur when the eye is fixated on a target. Tremors or nystagmus are rapid movements that occur during fixation. Because tremors likely have no functional relationship to vision, contemporary thought portrays tremors as noise created by the eye muscles (Spauschus, Marsden, Halliday, Rosenberg, & Brown, 1999). Vergence movements are used to shift the point of fixation from a close target to a distant one (divergence), or vice versa (convergence). The eyes rotate outward (in opposite directions) for divergence, and in towards each other for convergence. Although there has been

some investigation into vergence movements (e.g., Liversedge, White, Findlay, & Rayner, 2006; Yang, Bucci, & Kapoula, 2002), divergence and convergence are not frequently used as measures in typical laboratory procedures, because participants are normally held at a constant distance from relevant stimuli. “Smooth pursuit” refers to eye movements used to track an object moving through the visual field (Lindner & Ilg, 2006); smooth pursuit allows a moving target to remain in the fovea. Much like vergence, smooth pursuit movements are not relevant for the present eye tracking application as learning experiments tend to use static stimuli.

Eye tracking has proven to be valuable for its ability to provide a direct measure of overt attention during task performance. Eye movements have contributed massively in a number of clinical and industrial applications: from schizophrenia (Silverstein et al., 2015), autism (Guillon, Hadjikhani, Baduel, & Rogé, 2014), and Alzheimer’s disease (Fernández, Castro, Schumacher, & Agamennoni, 2015), to driving (Ahlstrom et al., 2013) and advertising (Higgins, Leinenger, & Rayner, 2014). Although eye movements have rarely been employed to directly investigate causality, eye tracking methodology has provided insights into attention to cues during learning (Beesley & Le Pelley, 2011; Kruschke, Kappenman, & Hetrick, 2005; Le Pelley, Beesley, & Griffiths, 2011, 2014), attention to context during learning (Lucke, Lachnit, Koenig, & Uengoer, 2013), and learned attention to drug related cues in addiction research (Hogarth, Dickinson, Austin, Brown, & Duka, 2008; Hogarth, Dickinson, & Duka, 2009, 2010a, 2010b; Hogarth, Dickinson, Hutton, Elbers, & Duka, 2006). Despite the utility of eye movements in studying attention in learning and a variety of other domains, the methodology has not gone without criticism.

The problem of covert attention. The eye-mind hypothesis assumes that the current point of fixation is the same as what is currently being processed (e.g., Just & Carpenter, 1980; Poole & Ball, 2006). However, it is difficult to deny that the relationship between oculomotor movements and attention can be problematic. Duchowski (2002) gives the example of using the

peripheral retina to look at stars. The light-sensitive rods, which are highly concentrated in the peripheral retina, allow for better localization of stars in the night sky than the cones of the fovea. Although the primary gaze point is elsewhere, attention is directed to the periphery. Using traditional eye tracking assumptions, it is difficult to accurately state where attention is directed because what is being processed is not in the fovea. People can direct gaze to location A, but direct attention to location B. In other words, covert attention can be a substantial problem when employing eye tracking, and research using eye tracking often comes under fire for this problem (Irwin, 2004). Klein (1980) observed that there was no difference in target detection latency regardless of whether the participant was asked to shift their gaze towards or away from the point they were cued to attend (see also, Klein & Pontefract, 1994). Trials in which overt and covert attention shifted in tandem had no benefit over trials in which attention was dissociated. Posner (1980) provides another critique for a strong connection between eye movements and attention, via evidence of attention shifts in the opposite direction of eye movement. Participants were able to execute two saccades (for a total of 16 degrees of movement), yet still reliably detect events back at the original fixation.

The evidence of dissociation leaves us with two distinct alternatives: (1) Gaze and attention are not related. People can, and frequently do disengage attention from the point of fixation, and (2) Gaze and attention are strongly linked. The loci of overt and covert attention frequently overlap. If the eye-mind hypothesis is invalid, then the conclusions of research using eye tracking as a measure of cognitive processes become significantly weaker. How can we ensure the eye-mind hypothesis is valid? Fortunately for the eye tracking researcher – despite the aforementioned reports of dissociation between eye movements and attention – a wealth of evidence supports a robust relation between eye movements and attention in many domains.

The non-problem of covert attention. Despite evidence supporting the problem of covert attention, the data frequently indicate that eye movements and attention are strongly

related, as opposed to the contrary. Hoffman and Subramaniam (1995) found that attentional shifts preceded saccades; their participants displayed difficulty in attending to a location while directing a saccade elsewhere, evidenced by decreased task performance when the saccade location and target location are incongruent. Deubel and Scheneider (Experiment 2; 1996) gave participants information on where a discrimination target would appear (via a cue), thus eliminating the possibility that participants could not find the target in time. The discrimination target was always presented in the same position of a stimulus array on the left or right. Despite knowing the exact location of the target stimulus, participants still showed enhanced performance when the discrimination and saccade targets coincided in location (see Deubel & Schneider, 2003 for similar findings). Griffin and Oppenheimer (2006) observed evidence of gaze at objects while preparing inaccurate, novel verbal labels for known objects (e.g., presented an image of a horse and saying the inaccurate label “blick”). The authors concluded that visual attention is automatically directed towards the object during processing, even when attending the object is unnecessary.

Studies integrating eye movements with human learning and reasoning also provide significant support for a relationship between attention and eye movements. Kruschke, Kappenman and Hetrick (2005) report one of the first applications of eye tracking for studying the allocation of attention during learning. Specifically, they examined gaze behavior during blocking (A-outcome, then AX-Outcome; diminished responding to X) and highlighting (AX-Outcome 1, then AY-Outcome 2; enhanced responding to Y). Results indicated that participants’ gaze behavior tracked cues that were predictive of outcomes (e.g. the highlighted cue), and diminished to cues that were non-predictive or redundant (e.g. the blocked cue) – although uninformative cues are never fully ignored. Le Pelley, Beesley and Griffiths (2011) employed an abstract predictive learning task in which nonsense words (e.g., conneastal, dusapplity, forditic) were predictive of sounds. Le Pelley et al., (2011) reported longer dwell time to stimuli that were

predictive of outcomes. In a second phase of the experiment the contingencies between cue and outcome were swapped, making the previously valid cues non-predictive, and the previously invalid cues predictive of outcomes. The attentional bias to the cues that were learned as predictive in the first phase was maintained even after the cue-outcome contingency changed, suggesting that attention to cues that aid learning and reasoning is preserved even after the cue's status changes, and that new learning is likely required to eliminate this attentional bias.

During learning and reasoning tasks, participants direct gaze to cues that are informative for making judgements or learning cue-outcome relationships. It is hard for people to ignore informative cues even when the cue's status changes. When cognitive resources are taxed – as they are in typical learning tasks – attention is more selective, meaning that dissociation of covert attention is less likely (Mitchell, Griffiths, Seetoo, & Lovibond, 2012; Quinlan, 2010). Rayner (2009) also suggests that task load results in strong overlap between overt and covert attention, and disengagement is due to planning the next saccade target; the dissociation between gaze and attention could be a "property of the processing system" (Rayner, 2009, p. 1458). Attentional dissociation is adaptive and can result from saccade programming, rather than a deliberate strategy used to ignore targets. In sum, attention and eye movements share a strong functional relationship. Given that task load ties overt and covert attention, that task performance suffers when covert attention is purposefully dissociated, and that gaze drifts towards objects of interest (even when visual attention is unnecessary), a dissociation between overt and covert attention is not likely to be a confound in the present studies.

Hypothesis

A great deal of the literature supports the idea that diagnostic reasoning differs from predictive reasoning because the order of diagnostic inferences is at odds with the order

engrained in our knowledge of causal relationships. One common hypothesis in *a priori* models is that, in order to make a diagnostic inference, people must re-map diagnostic relationships into predictive ones because the mental representation of the causal relationship is inherently predictive. This re-mapping would result in diagnostic inference-making requiring a greater degree of processing than predictive inference-making, and taking longer than predictive inference-making. Evidence of this remapping could be found in gaze behavior. A saccade/fixation pattern of switching gaze between the members of a causally-related pair of events would support the idea that participants are currently attending to the relationship between these two events. Finally, given that there is no inherent order of non-causal associations in *a priori* models, effects of order (on gaze or RT) should not be observed in forward or backward ordered non-causal pairs. Accordingly, for the present experiments, it was predicted that participants would spend more time gazing at diagnostic trials, complementing the longer reaction time in these same trials. More importantly, however, observations of alternations in gaze between the two members of the pair are expected, indicating that participants were actively processing the two concepts and the relationship between them. In addition to the reaction time difference, it was predicted that diagnostic trials will produce more alternations in gaze than predictive trials or any direction of non-causal association.

Chapter II: Gaze behavior during predictive and causal reasoning

Experiment 1

The purpose of Experiment 1 was to investigate differences in reaction time (RT) and gaze behavior when participants made predictive and diagnostic causal judgements. Fenker, Waldmann, and Holyoak (2005) observed shorter RT for predictive judgements compared to diagnostic judgements, and concluded that the difference was due to the greater difficulty of diagnostic judgements requiring more decision time. The RT difference is a consistent result, however, there is no data on what participants are doing during this extra decision time. From the *a priori* perspective, the longer judgement making time window in diagnostic inference should be due to re-mapping causal models. Experiment 1 is designed to investigate whether the difference in RT between predictive and diagnostic judgments was also present in gaze behavior. Specifically, do diagnostic judgments result in more gaze behavior (indicating more processing), in addition to the longer decision time? Experiment 1 employed a similar causal reasoning task as Fenker et al., (2005) in which participants made judgements on whether a pair of ideas (represented by a pair of words presented on a screen) shared a causal relationship or not. Under the assumptions of cognitive causal reasoning models, making causal inferences is effortful. Because gaze is frequently directed toward stimuli currently undergoing processing (in our case, two words on a screen) and participants' cognitive resources are taxed by the need to make repeated, accurate judgements, we expect participants to primarily direct both gaze and attention at the two members of the word pair. As previously stated in the hypotheses section, this gaze analysis is predicted to reveal re-mapping of diagnostic causal pairs as evidenced by alternations in gaze between the two members of the pair.

Method

Participants. For Experiment 1, 33 participants were recruited through Auburn University's online research service, SONA. SONA allows undergraduate students enrolled in psychology classes at Auburn University to participate in studies to earn extra credit points for their courses. All recruited participants were at least 19 years of age (or 18 with parental consent as per the law of the state of Alabama). All methods and experimental procedures were reviewed and approved by the Auburn University Institutional Review Board (IRB). Most studies of human causal learning recruit undergraduate students as participants. Although there has long been criticism of the use of convenience samples of undergraduates (e.g., Gallander Wintre, North, & Sugar, 2001), causal reasoning is a basic human cognitive process that does not seem to rely on social experience or socio-economic status. There is evidence that children engage in causal inference on a level similar to that of adults during late childhood (Bright & Feeney, 2014a; Fernbach, Macris, & Sobel, 2012). Thus, it is not likely that the undergraduate students sampled for the present studies display advanced or immature reasoning ability. Further, the gender of participants was not taken as a demographic variable. Investigating gender differences in causal inference is not currently a topic of debate in the causal learning literature. Given that causal inference is considered a building block of human cognition, an inherent assumption of most causality research is that effects of gender are minimal and do not confound results. This assumption is not without merit. Studies of childhood causal cognition typically recruit equally by gender, but do not find any gender differences (e.g., Gopnik et al., 2004; Sobel & Kirkham, 2006).

Materials. Participants completed the experiment in a room containing a single computer (Dell Precision T5400; Windows 7 OS) placed in a sound attenuating booth. The computer was connected to a Tobii T60XL eye tracker which sampled fixations at a rate of 60Hz. Participants sat roughly 18-20 inches from the eye tracker screen.

All word pairs were selected from Fenker et al., (2005). The authors normed selected words from the University of South Florida (USF) Word Association, Rhyme, and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998) to develop a list of causally and non-causally related pairs. The USF Word Norms list provides a large database of paired concepts that have been judged to be associated by a large sample of over 6,000 participants. For the researcher interested in obtaining large numbers of associated concepts, the database is an invaluable resource. Nelson et al., (1998) gave thousands of participants cue words, then had them write the first associated word that came to mind for each cue. Each word pair is cataloged with its forward (Word 1 → Word 2) and backward (Word 2 → Word 1) strength of association, the number of participants producing this association, and other metrics (e.g., number of other associates elicited by the cue word). All of the words Fenker et al., (2005) selected from this database had a low strength of association in both forward and backward directions – this was done to (a) prevent the salience of strongly associated pairs from affecting inference making RT and (b) ensure that all word pairs can be reversed without affecting inferences. Fenker et al., (2005) had a group of 80 participants judge whether there was a causal link between the words of an associated pair. If a causal link was present, participants gave a rating from 1-100 to the strength of that relationship; the presentation order of the words (Word 1 → Word 2 or Word 2 → Word 1) was counterbalanced across participants. The final word list created from the Fenker et al., (2005) norming study consisted of 68 word pairs that were judged to have a causal relationship, but did not significantly differ in predictive or diagnostic causal strength. Because of the rigorous vetting of the paired concepts for both association and causal strength, 40 causally related word pairs were selected from the Fenker et al., (2005) causal norming study for the two present experiments. Additionally, 40 pairs of non-causally related words were selected to use as comparison for order manipulation. As an example: “Tomato” → “Hamburger” are associated concepts, reversing the pair (to “Hamburger” → “Tomato”) does not affect a participant’s ability

to judge their association. The same holds for the predictive ordered concepts “Beat” → “Bruise”. (See Appendix A for a full list of words used in the experiment.)

Words were presented to participants via a task made in Tobii Studio 2.1. In the task, participants viewed a pair of words, and decided whether the two words were causally linked (i.e., one of the words is a cause of the other) or associated. The experiment was divided into two blocks. In each block, participants were given intermixed trials (in pseudo-random order) of associated and causally related pairs. The first block either consisted only of Word 1 → Word 2 ordered presentations (causal predictive and associated forward order), or Word 2 → Word 1 ordered presentations (causal diagnostic and associated backwards order). In the second block, a participant saw whichever order they had not yet seen.

Procedure. After entering the lab, and checking in with the experimenter on duty (to receive SONA credit and present evidence of parental consent, if necessary) participants sat in front of the eye tracker and went through the Tobii Studio calibration process. The calibration has the viewer gaze at nine distinct points on the screen (the four corners, four cardinal directions, and the center of the screen). If the first pass of calibration failed on at least one point, the experimenter on duty repeated the calibration on any failed points until the participant attained good calibration for all points. Next, participants read through the task instructions (See Appendix B for the task instructions), and went through eight practice trials to familiarize them with the task and response generation. Four of the practice trials presented Word 1 before Word 2, and four presented Word 2 before Word 1; all participants viewed the same pseudo-random sequence of trials. Half of the practice trials were causally related, and the other half were associated, but not causally related. The practice trials were presented via an E-prime 2.0 program (Psychology Software Tools, Pittsburgh, PA). On each trial, a fixation cross (white cross on black background; see Figure 5 for trial timeline) appeared on the left side of the screen for 1000ms. After the fixation cross disappeared, participants saw a black screen for

500ms, followed by the first word of the pair on the left side of the screen (white text on black background). After 1000ms of the first word alone, the second word appeared on the right side of the screen; both words remained on screen until the participant made a response: pressing “N" on the keyboard if the words were associated and not causally related, or pressing “C” on the keyboard if the words were causally related. After entering a response, participants received corrective feedback: either the word “Correct” in green text or “Incorrect” in red text presented on screen. The feedback remained on screen until the participant pressed the spacebar to continue. After the participant pressed the spacebar, they saw a black screen for 2000ms. After the eight practice trials, participants saw a screen to inform them that feedback was no longer available; pressing the spacebar on this screen began the test trials. Test trials were identical to practice trials with two exceptions. Firstly, the test trials were presented in Tobii Studio 2.1 rather than E-prime 2.0. Secondly, feedback was no longer available; thus, rather than pressing space to continue to the next trial, each subsequent trial automatically began 2000ms after entering a response. In the test trials, the 80 word pairs were separated into two blocks of 40 word pairs. Within each block, there were 20 associated word pairs and 20 causally-related word pairs. Each block had a predefined pseudorandom order of associated and causally-related pairs. One block contained the causal word pairs presented predictively and the associated pairs presented in the forward direction (Predictive Block). One block contained the causal word pairs presented diagnostically and the associated pairs presented in the backwards direction (Diagnostic Block). Participants were randomly assigned to either see the Predictive Block or the Diagnostic Block first. After completing all 80 test trials, participants saw an end screen with a message thanking them for their participation in the study and were dismissed from the lab.

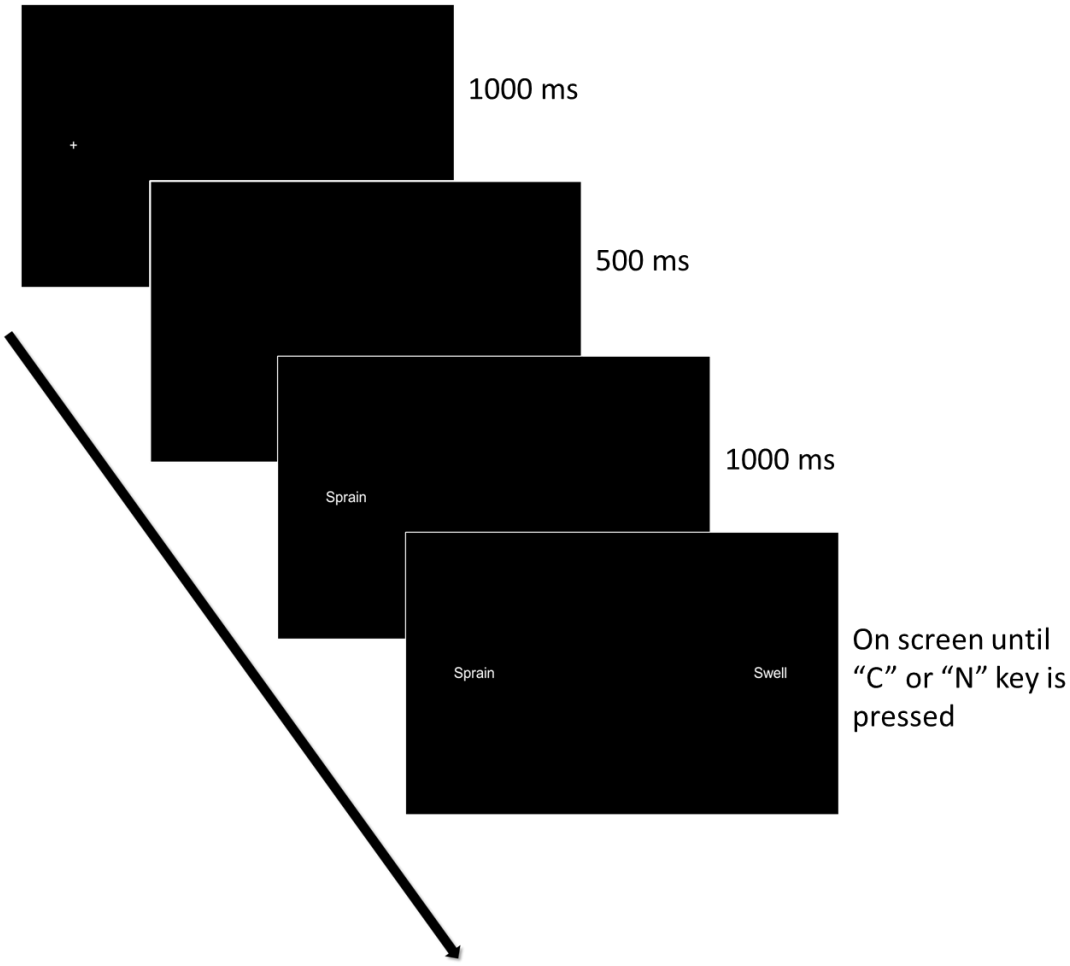

Figure 5. Trial timeline for Experiment 1.

Results and Discussion

The data were analyzed with a 2x2x2 (Causal Order x Relationship Type x Block) repeated measures ANOVA (see Appendix C). Participants were placed in the predictive-first group or diagnostic-first group depending on which block they completed first. The key dependent variables were: reaction time (measured as the time between the moment when Word2 is presented, and the moment when the participant actually makes a response), accuracy in categorizing each trial, and alternations in gaze between Word1 and Word2

obtained from the eye tracker data (again measured between response availability and the actual response). Raw reaction time data was normalized using a base 10 logarithm transformation. In the eye tracking software Tobii Studio, two areas of interest (AOIs) corresponding to the position of the two words were created for gaze analysis: AOI1 on the left side of the screen, and AOI2 on the right side of the screen. A shift in gaze from AOI1 to AOI2 is counted as one alternation. One participant performed below 50% accuracy across trials and therefore, the participant's data was not included in the subsequent analyses.

Firstly, the three-way interaction of Block, Relationship Type, and Causal Order was significant ($F(1,29) = 6.56$, $p = .02$, partial $\eta^2 = .18$). Splitting the data by Block (Figure 6) reveals an interaction between Relationship Type x Causal Order ($F(1,29) = 11.90$, $p = .01$, partial $\eta^2 = .29$). Using paired-samples t-tests for post-hoc analysis indicated that these interactions were driven by differences in the predictive-first group, and that there were no significant differences between the trial types in the diagnostic-first group. In the predictive-first group, there were significant differences between causal predictive and causal diagnostic trials ($t(17) = 5.43$, $p < .001$). Causal predictive trials significantly differed from forwards ($t(17) = 2.51$, $p = .02$), but not backwards associated trials ($t(17) = 1.17$, $p = .26$). Causal diagnostic trials significantly differed from both forwards ($t(17) = 3.29$, $p = .004$) and backwards associated trials ($t(17) = 4.37$, $p < .001$). That is, predictive-first participants took significantly longer to categorize causal diagnostic word pairs ($M = 2732$, $SD = 674$) compared to predictive causal ($M = 2066$, $SD = 536$) and associated trials ($M = 2310$, $SD = 653$; $M = 2201$, $SD = 577$). The data of the predictive-first group are consistent with the asymmetry observed by Fenker et al., (2005) in that order manipulation had a greater effect on causally-related pairs than on associated pairs as association judgements do not rely on causal structure. The data of the diagnostic-first participants displays no such asymmetry.

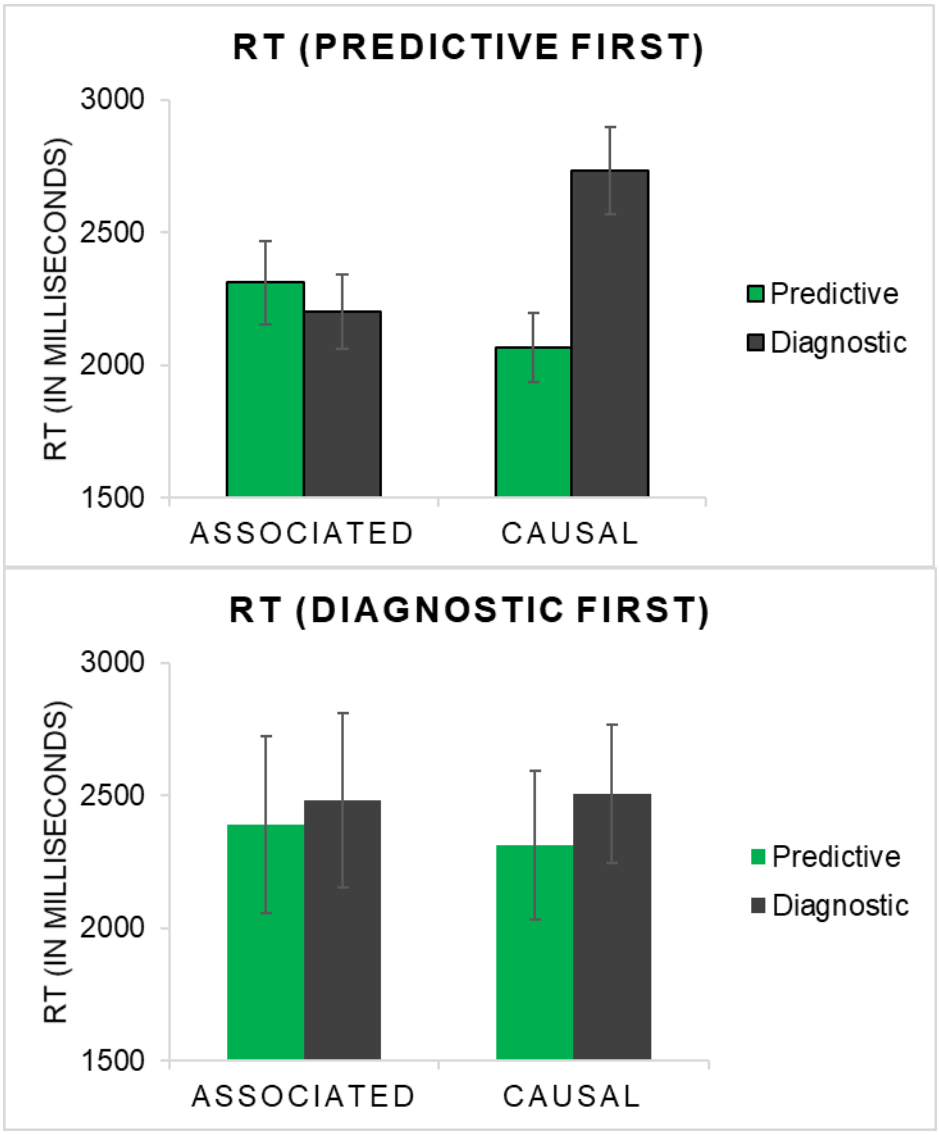

Figure 6. Reaction time (RT) data from Experiment 1. Error Bars are the standard error of the mean (*SEM)*.

Next, participants' accuracy (Figure 7) at categorizing trials was examined using a 2x2x2 repeated measures ANOVA. There was no significant three-way interaction between Block, Relationship Type, and Causal Order ($F(1,29) = 1.49$, $p = .23$, partial $\eta^2 = .05$). However there was a significant interaction between Relationship Type and Causal Order ($F(1,29) = 5.04$, $p = .03$, partial $\eta^2 = .15$). In the absence of any other significant Block interactions (See Table 5), the

following analyses collapse across group. Post-hoc analyses (paired-samples t-tests) indicated that there was a significant difference in average accuracy ($t(31) = 2.70$, $p = .01$) between predictive causal ($M = 0.80$, $SD = 0.08$) and forwards associated trials ($M = 0.86$, $SD = 0.11$) but not ($t(31) = .97$, $p = .34$) between predictive causal and backwards associated trials ($M = 0.82$, $SD = 0.13$). Further, the average accuracy for diagnostic causal trials ($M = 0.67$, $SD = 0.19$) was significantly lower than average accuracy for both predictive causal ($t(31) = 3.93$, $p < .001$) and both directions of associated trials ($t(31) = 4.37$, $p < .001$; $t(31) = 3.44$, $p = .002$). As hypothesized, performance on diagnostic trials was once again different from predictive and associated trials; diagnostic trial accuracy was significantly worse than the other trial types. This dip in accuracy can be interpreted as evidence of the purported difficulty of diagnostic causal reasoning. Causal order thus seems to be deterministic for accuracy. Both findings so far provide some evidence that supports the predictions of models prioritizing *a priori* causal structure.

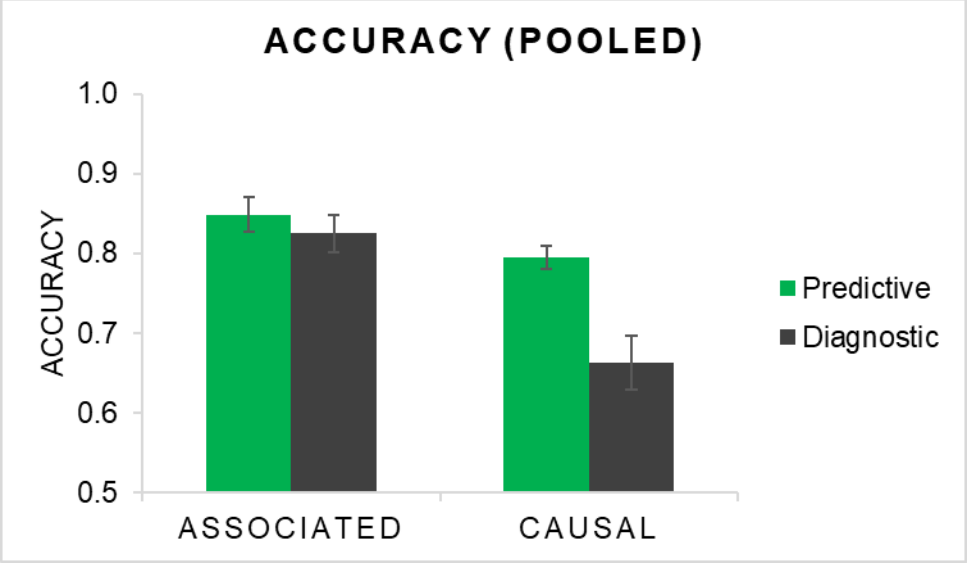

Figure 7. Accuracy data from Experiment 1. Error Bars are *SEM*.

Finally, average gaze alternations were analyzed using a 2x2x2 repeated-measures ANOVA (See Table 6). There was no significant three-way interaction between Block, Relationship Type, Causal Order ($F(1,29) = .08$, $p = .78$, partial $\eta^2 = .003$). However, Block did interact with Causal Order ($F(1,29) = 6.87$, $p = .01$, partial $\eta^2 = .19$) Once again, Relationship Type interacted with Causal Order ($F(1,29) = 4.74$, $p = .04$, partial $\eta^2 = .14$). Paired-samples t-tests indicated that, for diagnostic-first participants, there were no significant differences among the trials types. For predictive-first participants, predictive causal pairs of words ($M = 0.92$, $SD = .78$) and forwards associated pairs of words ($M = .95$, $SD = .83$) did not significantly differ ($t(17) = .39$, $p = .7$) in average gaze alternations, and backwards associated pairs ($M = 1.09$, $SD = .97$) did not significantly differ from predictive causal pairs ($t(17) = 2.00$, $p = .06$). The difference between trial types in predictive first was driven by the diagnostic trials ($M = 1.31$, $SD = 1.07$). Specifically, diagnostic trials produced significantly more alternations than predictive causal ($t17) = 3.44$, $p = .003$), forwards associated ($t(17) = 3.55$, $p = .002$) and backwards associated ($t(17) = 2.23$, $p = .04$) trials.

The predictive-first alternation data serves to complement the predictive-first RT and accuracy data. Diagnostic causal reasoning not only takes longer and is more difficult than predictive causal reasoning and non-causal reasoning, but this increased RT is invested in more alternating (and thus, processing) between the two members of the word pair. The diagnostic-first data seems to support a different conclusion. Diagnostic-first produces attenuation of the expected asymmetry of causal order. Some potential reasons for this effect will be addressed in the General Discussion. Accordingly, the results of Experiment 1 demonstrate partial support for asymmetry of causal order – predictive inference differs from diagnostic inference in speed, accuracy, and processing features when there are no attenuating manipulations. Causal judgements were also demonstrated to be sensitive to order whereas non-causal judgements were order-agnostic. In other words, Experiment 1 provides some support for the predictions of

a priori models of causal learning and inference, and provides novel evidence that gaze behavior reveals the re-mapping of diagnostic causal relationships.

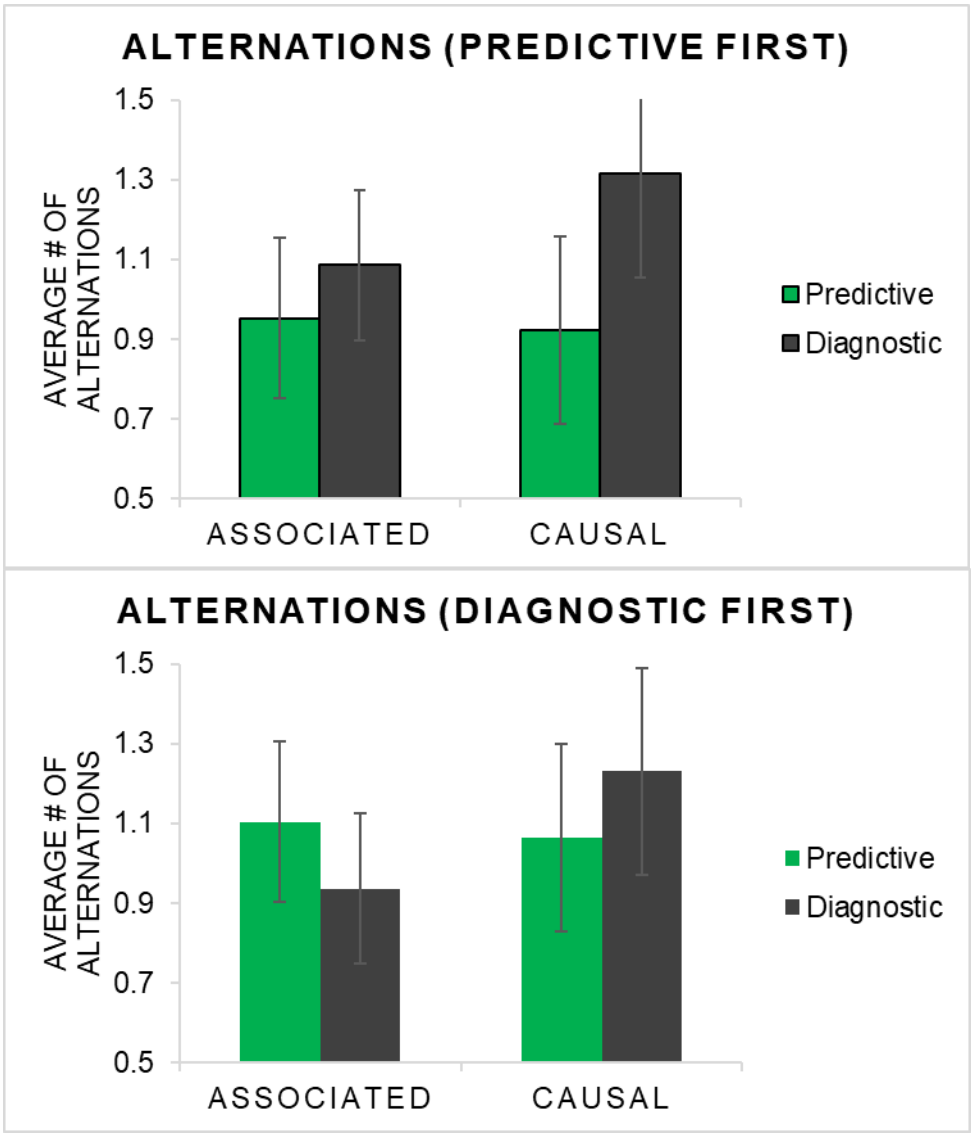

Figure 8. Gaze data (alternations between first and second word of the pair) from Experiment 1. Error Bars are *SEM*.

Experiment 2

Experiment 1 replicated the RT effect of Fenker et al., (2005) in which people are slower in making diagnostic judgements than predictive judgements. Secondly, Experiment 1 provides evidence that differences between predictive and diagnostic reasoning are also observable in alternations of gaze. One potential confound in Experiment 1 was that participants could have used peripheral vision to maintain continuous access to both of the words presented in each trial. Nevertheless, the difference in average gaze alternations provides evidence of differential processing in predictive and diagnostic causal reasoning. Furthermore, the potential effect of peripheral vision would equally affect the processing of all three types of word pairs. Gaze alternations, rather than peripheral vision, allow a participant continuous access to the pair of words in the trial while they make an inference. However, how will the processing of causally-related pairs of words be affected if we restrict the participants' ability to alternate, and thus the continuous access to both members of the word pair? The purpose of Experiment 2 was to eliminate the participants' continuous access to the pair of words by replacing the first word of the pair with a placeholder after the presentation of the second word. People seem to gaze at stimuli that serve as referents of what is currently being processed, regardless of the actual content of said stimulus. Kahneman and Lass (as cited in Just & Carpenter, 1976) reported that participants would gaze at symbols that referred to the category in which they were answering questions. For example, when naming types of cars, participants would spend the most time looking at the image of a car. Gaze was also directed to the former location of the image when the actual image has been removed. Griffin and Oppenheimer (2006) observed evidence of gaze at images of objects while preparing inaccurate, novel verbal labels for known objects. Taken together, these studies suggest that that visual attention is often directed towards the stimulus being processed, even when attending the stimulus is unnecessary. Thus, a hypothesis of Experiment 2 is that participants, although forced by the experimental design to

focus only on Word2, would initially alternate between the placeholder of Word1, and the actually presented Word2. Due to the manipulation of Experiment 2, the alternations will not provide access to the pair of words during processing, accordingly, gaze alternations are not expected to be influenced by the difficulty of the trial. Thus, the average number of alternations on diagnostic causal trials will not differ when compared to predictive causal trials or non-causal trials. As in Experiment 1, it is expected that the accuracy of participants' judgements will be higher for the predictive and associative pair of words, and lower for the diagnostic trials. In a similar fashion, it is expected that the RT for the diagnostic trials will be significantly longer than the RT for the predictive and associative trials.

Method

Participants. For Experiment 2, 23 participants at Oakland University in Rochester, MI were recruited. To maintain consistency between Experiments 1 and 2, all participants in Experiment 2 were at least 18 years of age. All methods and experimental procedures were reviewed and approved by the Oakland University Institutional Review Board (IRB).

Materials. The same causally-related and associated words from the USF Word Norm Association list (Nelson et al., 1998) selected for Experiment 1, were used in Experiment 2. (See Appendix A for a list of words used in the experiments.)

Once again, words were presented to participants via a task made in Tobii Studio 2.1. In the task, participants viewed a pair of words, and decided whether the two words were causally linked or associated. As in Experiment 1, participants saw causal trials in both predictive and diagnostic causal order, and saw associated trials in both forward and backward order.

Procedure. Participants began by checking in with the experimenter on duty, then sat in front of the eye tracker to go through the calibration process. Next, participants read through the task instructions (See Appendix B for the task instructions) and completed eight practice trials to

familiarize them with the task and response generation. The practice trials, created in E-Prime 2.0, were identical to the trials used in Experiment 1. After completing the practice trials, participants began the experimental task. At the beginning of each test trial, participants viewed a white fixation cross on the left side of the screen (in the center of the Word1 location) on a black background. After 1000ms, the cross disappeared and an empty black screen appeared for 500ms. The first word of the pair then appeared on the left side of the screen for 1000ms, followed by a screen including both the first word and the second word (which appeared on the right side of the screen). Both words remained on screen for 1000ms, then the first word was replaced by a foil – a pseudo-word of equivalent length with the same first letter. For example, the foil of the word “Hamburger” would be “Hxxxxxxxx” (see Figure 9 for a timeline of the Experiment 2 task). The foil and Word2 remained on screen until the participant made a response. After each trial, a black screen appeared for 2000ms.

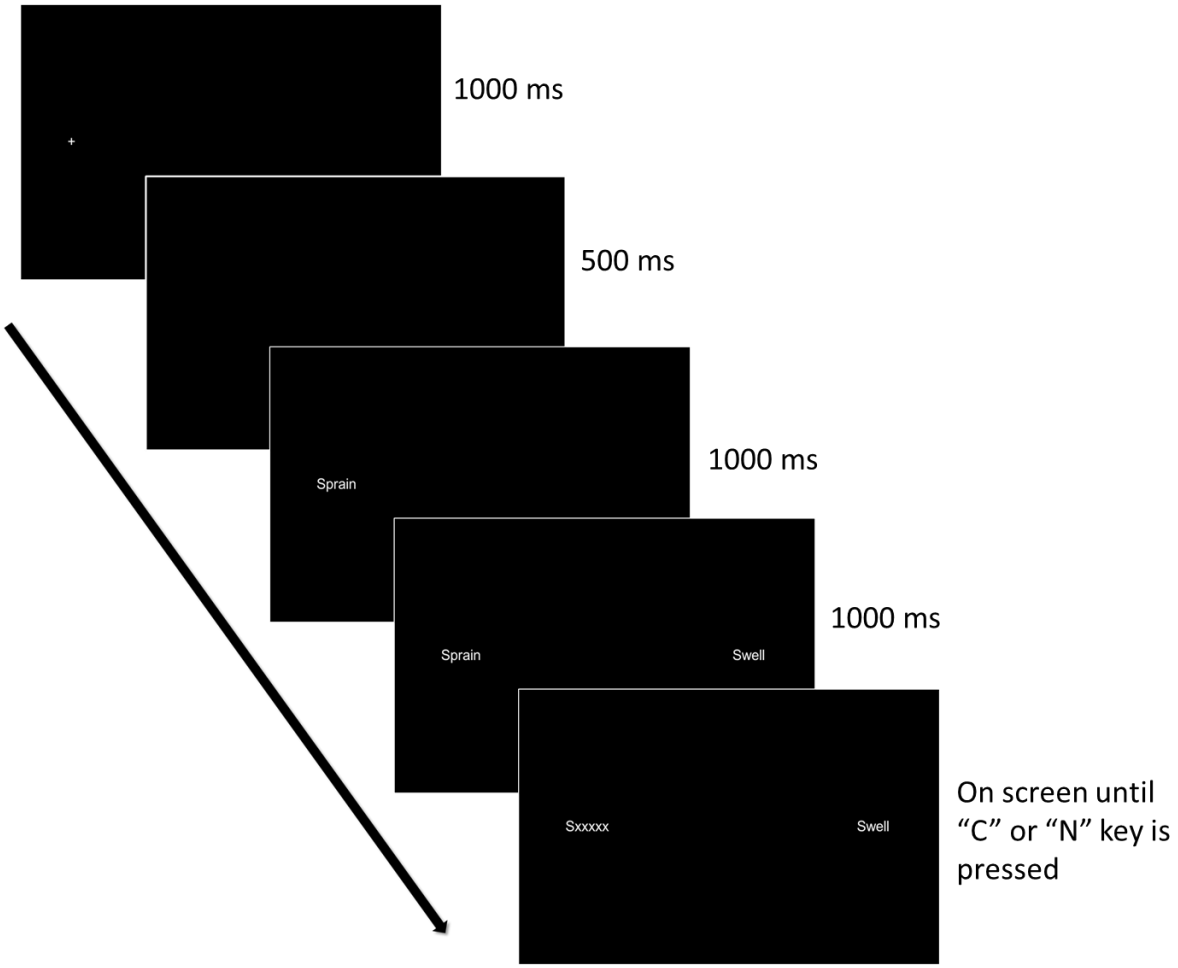

Figure 9. Trial timeline for Experiment 2

As in Experiment 1, the words pairs were separated into two blocks of 40 word pairs, one predictive block and one diagnostic block. Participants were randomly assigned to either see the predictive or diagnostic block first. Each block contained 20 associated word pairs and 20 causally-related word pairs. The predictive block contained only forward ordered associated pairs, and the diagnostic block contained only backwards ordered associated pairs. Each block had a predefined pseudorandom order of associated and causally-related pairs. After completing all 80 test trials, participants viewed an end screen with a message thanking them for their participation and were dismissed from the lab.

Results and Discussion

As in Experiment 1, these data were analyzed using a 2x2x2 (Causal Order x Relationship Type x Block Order) repeated measures ANOVA. Raw reaction time data was again normalized using a base 10 logarithm transformation. Two areas of interest (AOIs) were defined in Tobii Studio for gaze analysis. Some participants ended up with insufficient available gaze data for significant portions of the experiment. Trials are very sparse in terms of gaze behavior when the valid gaze percentage is low, making it difficult to identify alternations. Low valid gaze percentage is often due to participants looking away from the screen for a significant amount of time. Five participants with percentages of valid gaze points at or below 50% were dropped from the analyses discussed below. Thus, only the data from 18 participants was used. Dependent variables were reaction time, accuracy, and alternations between Word1 and Word2. RT and alternations are measured between the moment a response becomes available and the moment a participant actually responds.

For RT (Figure 10), no significant three-way interaction was found between Block, Relationship Type, and Causal Order ($F(1,16) = .25$. $p = .63$, partial $\eta^2 = .02$), thus the following analyses are collapsed across Block. There were no other significant interactions or main effects (See Table 7). The moderate effect size indicates that the lack of significant effects may be driven by the small sample size. The large standard deviations for these means also point towards the increased variability in the present experiment. Looking at the means of the three trial types reveal trends that are, on the surface, consistent with predicted results. The RT for forward (M = 1,821, SD = 1196) and backwards (M = 1,854, SD = 1109) associated trials differ from predictive causal trials (M = 1,816, SD = 1071) differ by 5ms and 38ms respectively. In contrast, there was a difference of 335ms between causal diagnostic trials (M = 2,151, SD = 1395) and causal predictive trials, and differences of 330ms and 297ms between causal diagnostic and associated trials.

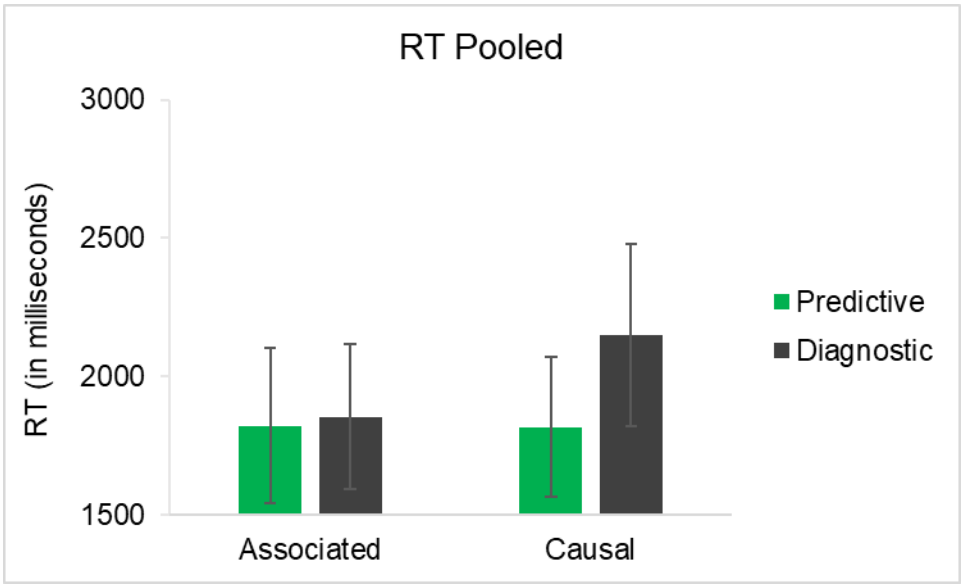

Figure 10. Reaction time (RT) data from Experiment 2. Error Bars are *SEM*.

No significant three-way interaction was found between Block, Relationship Type, and Causal Order ($F(1,16) = .1.52$, $p = .24$, partial $\eta^2 = .09$), thus the following analyses of average accuracy are collapsed across Block. Block did not significantly interact with Relationship Type or Causal Order (See Table 8). The only significant main effect was that of Causal Order ($F(1,16) = 11.39$, $p = .003$, partial $\eta^2 = .42$). The results of a post-hoc test (paired-samples t-test), indicate a significant difference between predictive and diagnostic trials ($t(18) = 2.11$, $p = .049$). Participants performed worse on diagnostic trials ($M = 0.69$, SD = 0.158) than on predictive causal trials ($M = 0.77$, $SD = 0.112$) The performance on predictive causal and diagnostic causal trials appears similar to that of Experiment 1. However, associated trial performance seems significantly worse than in Experiment 1. Based on prior research and theory, lower accuracy due to the increased difficulty of diagnostic causal reasoning would be expected. Associated trial performance was likely affected by the absence of Word2 (located on

the left side or the screen in AOI1) when participants made a judgement. The manipulation made it necessary to maintain the first word in memory during judgement making during all trial types. Thus, it is not clear why the absence of one of the two words did not affect accuracy to the same degree in predictive causal trials.

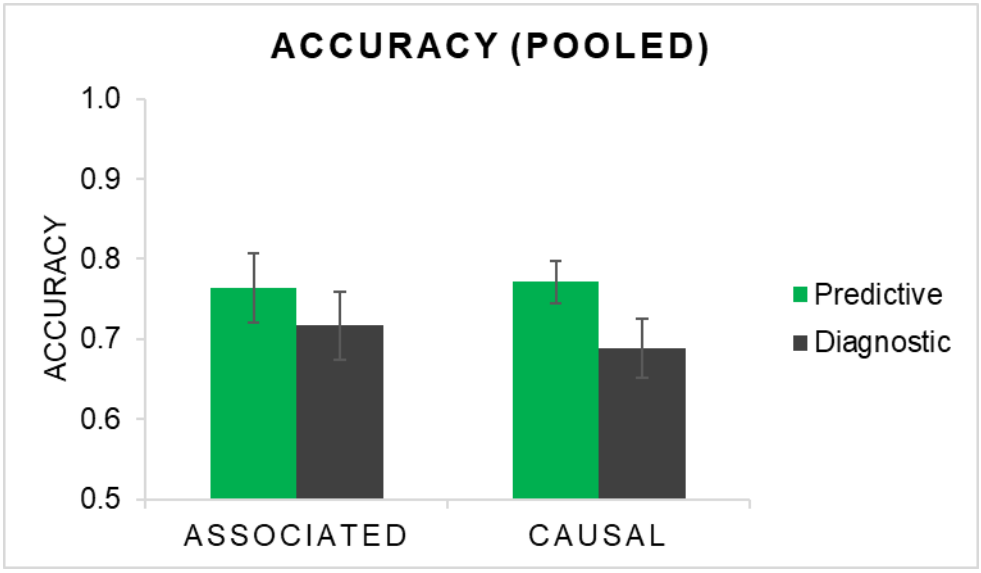

Figure 11. Accuracy data from Experiment 2. Error Bars are *SEM*.

For number of alternations (Figure 12), there was no significant interaction between Block, Relationship Type, and Causal Order ($F(1, 16) = 4.07$, $p = .06$, partial $\eta^2 = 0.20$). There were no other significant interactions or main effects (See Table 9). In general, there were slightly more alternations in causal predictive ($M = 1.09$, $SD = 1.14$) and causal diagnostic ($M = 0.96$, $SD = 1.21$) trials than in forward ($M = 0.92$, $SD = .90$) and backward associated trials (M = 0.79, SD = 0.77). The absence of a word in the antecedent location during decision making makes it less likely that participants will go back and look at the placeholder. Accordingly, all conditions suffered from low numbers of alternations (on average less than one per trial). The

presence of alternations at all lends some credence to the idea that people still direct gaze at what is being processed even when it is not necessary for judgement making.

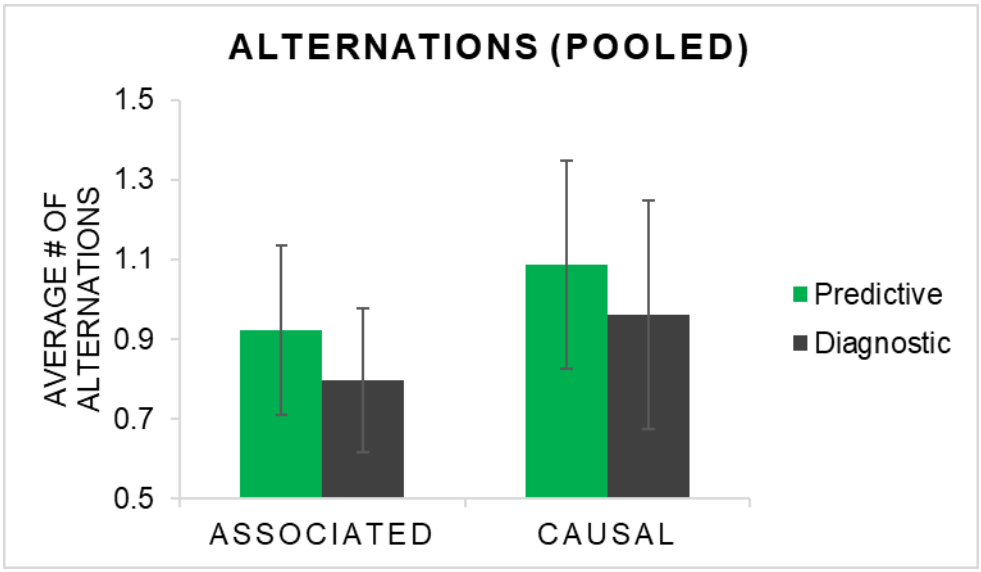

Figure 12. Gaze data (alternations between first and second word of the pair) from Experiment 2. Error Bars are *SEM*.

Among participants with valid gaze points above the 50% criteria, the diminished gaze behavior may have been a learned response during the course of the experiment. The first word in the pair was always obscured during the response period. It is possible that participants began the experiment allocating attention to AOI1, but later learned to direct attention away from AOI1 and primarily focus on AOI2, as AOI1 does not provide any relevant information during the response period of test trials. To examine this learned inattention hypothesis, the data were split into blocks of five trials to examine whether there was any change in number of alternations during the first few test trials compared to the last few test trials. There were no significant main effects or interactions in the 2x2x4 (Relationship x Direction x Block#) repeated measures ANOVA. Thus, the average number of alternations did not change across block of trials. However, it is notable that the main effect of block has a moderate effect size (partial η^2 =

0.10). As a second attempt to find evidence of a learning effect reducing gaze behavior, the average duration of gaze to the antecedent word (AOI1) and to and subsequent word (AOI2) in each trial (fixation duration by AOI) was calculated. Again, there were no significant main effects or interactions in the 2x2x2 (Relationship x Direction x AOI) repeated measures ANOVA, although the main effect of AOI did have a notable effect size (partial $\eta^2 = 0.15$).

Finally, the results of Experiments 1 and 2 were compared using a three-way (Relationship Type, Causal Order, and Experiment) repeated measures ANOVA (See Table 10, Table 11, and Table 12). For reaction time, there was a significant interaction between Relationship Type and Causal Order ($F(1,47) = 7.12.$, $p = .01$, partial $\eta^2 = 0.13$), however, there was no Relationship Type x Causal Order x Experiment interaction (p = .22). Similarly for accuracy, there was a significant Relationship Type x Causal Order interaction ($F(1,47) = 4.06$, $p = .49$, partial $\eta^2 = 0.08$), but no significant three-way interaction ($p = .33$). For alternations, there was no significant Relationship Type x Causal Order interaction ($p = .19$) or three-way interaction ($p = .19$). However, Causal Order did interact with Experiment ($F(1,47) = 6.20$, $p = .02$, partial $\eta^2 = 0.12$). Post-hoc analyses (independent samples t-tests) for RT, indicated that reaction times were significantly faster across all trial types in Experiment 2 (see Table 13). For accuracy, only backwards associated trials significantly differed between the two experiments ($t(47) = 2.36$, $p = .02$). For average alternations there were no significant differences between the means ($ps > .05$). (See Appendix D for experiment comparison graphs.) The Relationship Type x Causal Order interaction in these analyses merely reflects the effect of diagnostic presentation on causal inference uncovered in Experiments 1 and 2 and leads to no additional conclusions. These analyses indicate that the masking of the second word in Experiment 2 had no significant effect on participants' accuracy in categorizing a trial or the pattern of their gaze behavior. When it comes to gaze behavior, the mean for diagnostic trials has the largest difference across the two experiments (although the difference in means was not significant).

This difference suggests that the mask affects alternations on diagnostic trials more than other trial types. The mask however, does seem to result in participants completing trials significantly faster than when no mask is present. This RT effect of masking Word 2 will be elaborated on in the General Discussion.

The data of Experiment 2 show slight trends towards the conclusions drawn from prior research, however, it seems that the low *N* in this study constrains most attempts to explain the pattern of results. A tentative explanation could be that participants may have alternated less at the end of the experiment than at the beginning, but power to detect this difference was insufficient.

General Discussion

Many modern theories of causality prioritize the role of *a priori* causal structure, and portray the processes of causal learning and reasoning as cognitively taxing, sensitive to manipulating causal order (and other cues to causal structure), and frequently non-normative. Currently, there is no strong consensus on the best way to model how people come to the conclusion that there is a causal connection between two events. However, the literature makes it clear that: (a) Covariation alone is insufficient to account for causal learning – the asymmetry between prediction and diagnosis (as observed in the present studies) cannot be explained through covariation. (b) A distinction between causal and non-causal events is integral to the human cognitive system – asymmetry is not present when the events are not causally related. (c) The structure of a relationship has importance for causal induction – it makes intuitive sense for knowledge of a causal relationship to play as large a role as (or a larger role than) the actual parameter describing the strength of the relationship. The results of Experiment 1 provide some evidence that fits with this cognitive perspective of causality. The observed asymmetries between assessing predictive and diagnostic causal relationships are absent in assessing forward and backward ordered non-causal relationships. The present studies support the idea

that causal reasoning, in general, is cognitively effortful, and that diagnostic causal reasoning requires more effortful processing than predictive causal reasoning or non-causal reasoning. Specifically, diagnostic causal reasoning takes longer and produces lower accuracy than both predictive causal reasoning and non-causal reasoning. More importantly, the longer decision making time seems to be associated with more effortful processing by participants, as evidenced by the larger number of gaze alternations in diagnostic causal trials. In Experiment 1 these findings are modulated by effects of training. When diagnostic trials are practiced, as in the data of the diagnostic-first participants, these causal order effects are absent. In the predictive-first data no such attenuation occurs. The data of Experiment 2 also illustrate trends resembling previous research. The RT data contains similarities to both Experiments 1 and the Fenker et al., (2005) data, in that causal diagnostic trials take the longest to categorize. However, the single word paradigm of Experiment 2 creates significantly faster reaction times among participants compared to the dual word paradigm in all trial types. Further, accuracy is similar to Experiment 1, in that causal diagnostic trials have the lowest accuracy. These data of Experiment 2 are consistent with the ideas behind a cognitive approach to understanding causality. Specifically, that diagnostic reasoning is more difficult than predictive reasoning due to re-mapping the information to fit the causal model, and that the increased difficulty in diagnostic reasoning results in longer decision making time than predictive or non-causal reasoning. The alternation data of Experiment 2, does not match up to that of Experiment 1 – partially due to dropped participants having a significant impact on the ability to detect any gaze effects. Participants gradually learning NOT to regress to the location of the antecedent word (AOI1) is an intriguing possibility ultimately unsupported by the data. The alternation data does indicate that, as hypothesized, participants do direct some gaze to the non-informative AOI1. Taken together with the RT data, participants engage in less processing when they only have access to a single member of the pair, and the diminished processing is reflected via shorter

overall decision making time and reduced gaze behavior. The visual mask seems to change the task by fundamentally altering what participants are being asked to do. When participants are not able to alternate they do not have continuous access to the relationship; it seems that they must then rely on their memory of Word1 and its relation to Word2. Instead of making a judgement about the two words on screen, participants may have to access the mental representation of the causal relationship in order to make a judgement. In this case, no alternating between the stimuli or remapping would be required to succeed at the task. It is also possible that the masking of Word1 increased the cognitive load elicited by the task. Not having access to the words they are currently judging adds an extra dimension to the task at hand. Similarly, Bright and Feeney (2014b) observed attenuation of causal asymmetry caused by increased cognitive load.

Limitations

One of the major limitations of Experiment 1 was the use of a block design (presenting all predictive trials, then all diagnostic trials, or vice versa) for comparison of predictive and diagnostic trials. Although the block design was selected for the ability to compare to past studies, unexpected effects of block order emerged. As previously stated, Block One biases the subsequent block, either enhancing or attenuating asymmetry of causal order. An alternate method of designing the studies would have been to intermix the three trial types (predictive causal, diagnostic causal, associated) pseudo-randomly to prevent runs of a single trial type. This limits the ability of participants to commit to a single strategy for judgement making, and eliminates effects from participants being forced to switch strategies halfway through the experiment. Although the effects of block order may be interesting to investigate directly, using an intermixed design for predictive and diagnostic judgements would make for a stronger test of the re-mapping proposed by *a priori* models. Experiment Two was not affected by the block design, however, this may be due to the reduced sample size in the experiment.

Experiment Two was plagued by its own set of issues. In order to address the limitations of Experiment 2, the following ideas could prove fruitful. In the future, valid gaze point percentage should be a broad exclusion criterion. Rather than simply collect a set number of participants (e.g., 30), a set number of participants who meet the criterion should be collected. Participants with less than 50% gaze point validity add little to no useful gaze data, as low gaze point validity is a function of participants' performance. Thus, keeping only the data of participants with consistent gaze data could solve the problems of high variability and reduced power that occurred in Experiment 2. Secondly, as previously stated, the visual mask fundamentally changes the nature of the task. Accordingly, for a purer test of the gaze effects, adding a condition in which Word1 returns to the screen after the foil is briefly presented could be used to assess whether participants are learning to ignore a low information gaze point (AOI1). Participants learning to ignore Word1 would be a finding that fits with trends in the literature. Research into attention allocation during learning (Beesley & Le Pelley, 2011; Le Pelley et al., 2014; Mackintosh, 1975) suggests that people learn to direct attention away from non-informative stimuli (but see Kruschke & Blair, 2000; Kruschke et al., 2005 for an alternate view). In addition to these design flaws, there were potential issues with uncontrolled variables in both the stimuli and the sample.

The stimuli from the USF Word Norm list were controlled for strength of association and causal strength in both forward and backwards directions. However, certain features of the word pairs were not controlled. The length of the words in each pair were not equivalent. There were some very long words (e.g., Compliment, Employment, Protestants) and some words with half as many characters (e.g., Rush, Acid, Scar). It is possible that these longer words take longer to parse than shorter words, and this could have produced effects in RT or gaze behavior on those trials. Similarly, the frequency of words was uncontrolled. Some words and word pairs are very common in everyday language (e.g., Dog-Claw, Attack-Defense) while others were more

obscure (Lime-Corona, Caffeine - Mountain). The more obscure pairs might also contribute to difficulty of specific trials. For future studies employing similar paradigms, a more uniform stimulus set would solve these problems. For instance, norming words for usage frequency, and specifying a range of character lengths for every word in the set.

Besides race, age, and gender, there are other, more dangerous, uncontrolled factors present within the sample. Firstly, given that the task requires participants to make judgements based on words they read, some degree of verbal ability is recruited. Accordingly, any differences in the reading/verbal ability of the sample may also be reflected in the between-subjects differences in the data. Individual differences in reaction time likely had a similar effect on the data. Response time in the present studies required perceiving the stimuli, processing the causal relationship, and then finally generating the mechanical response. Participants differ in their ability to perform all of these stages. Collecting some measure of individual RT, as well as a measure of verbal ability would provide valuable data to researchers in the future.

Future directions

Investigators interested in causality can take future research incorporating gaze behavior and causal inference in many interesting directions. Block order produced some unexpected results in Experiment 1. Future studies could deliberately examine the attenuating effects of making diagnostic inferences prior to making predictive inferences. Specifically, the questions of interest are: Firstly, whether the attenuation of causal asymmetry is a replicable effect. Secondly, does the inverse, making predictive inferences prior to making diagnostic inferences, result in enhancement of causal order asymmetry?

In Experiment 2, the visual mask forced participants to rely on memory for Word 1 when making a judgement. Accordingly, future research could directly compare memory for causal relationships to visually presented causal information. If the present results are replicated, then we would expect significantly lower RT for memory. When participants are relying on memory, *a*

priori models propose that they are relying on an already constructed (predictive ordered) causal map, thus no remapping is required for diagnostic inference making. Memory could also be manipulated by replacing both words with irrelevant foils (or images representing the causal pair). Participants will still have locations to alternate between, but no relevant information. Gaze behavior and reaction time would then provide insight into whether participants are employing causal maps.

Finally, future research could employ use a secondary (non-visual) task to tax participants' cognitive resources while measuring gaze behavior. Prior research has found that increasing cognitive load attenuates asymmetry of causal order (e.g., Bright and Feeney, 2014b). The mask used in Experiment 2 likely increases cognitive load, but also encourage participants to visually disengage from the task. Using a secondary task during inference making might allow for a cleaner investigation of this hypothesis. A secondary task would depress gaze behavior while avoiding disengagement. However, the secondary task may also increase RT rather than decrease RT, as did the masking in Experiment 2.

Conclusion

The study of causality has produced answers to many of the questions that we have about the lens that shapes our perception of the world. Asymmetry of causal order is a strong prediction of models of causality that propose the use of *a priori* causal structure to build mental representations of causal relationships. The asymmetry present in behavior is the result of covert mental processes. The present experiments sought to give insight into these covert processes. Preliminary evidence of re-mapping during diagnostic inference was observed. However, additional investigation is needed to elucidate the conditions under which people engage in re-mapping and to attune causal reasoning tasks to detect the differences between predictive and diagnostic inference.

References

Ahlstrom, C., Nyström, M., Holmqvist, K., Fors, C., Sandberg, D., Anund, A., … AAkerstedt, T. (2013). Fit-for-duty test for estimation of drivers' sleepiness level: eye movements improve the sleep/wake predictor. *Transportation Research Part C: Emerging Technologies*, *26*, 20–32.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye Movements Do Not Reflect Retrieval Processes Limits of the Eye-Mind Hypothesis. *Psychological Science*, *15*(4), 225–231.

Arcediano, F., Matute, H., Escobar, M., & Miller, R. R. (2005). Competition between antecedent and between subsequent stimuli in causal judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 228.

Barnes, J. (1991). *The Complete Works of Aristotle, The Revised Oxford Translation, ed. Jonathan Barnes*. New Jersey: Princeton University Press.

Beesley, T., & Le Pelley, M. E. (2011). The influence of blocking on overt attention and associability in human learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 114.

Bright, A. K., & Feeney, A. (2014a). Causal knowledge and the development of inductive reasoning. *Journal of Experimental Child Psychology*, *122*, 48–61.

Bright, A. K., & Feeney, A. (2014b). The engine of thought is a hybrid: Roles of associative and structured knowledge in reasoning. *Journal of Experimental Psychology: General*, *143*(6), 2082.

Castelhano, M. S., & Rayner, K. (2008). Eye movements during reading, visual search, and scene perception: An overview. *Cognitive and Cultural Influences on Eye Movements*, 175–195.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 837.

Cobos, P. L., López, F. J., Cano, A., Almaraz, J., & Shanks, D. R. (2002). Mechanisms of predictive and diagnostic causal induction. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(4), 331.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1–20.

De Houwer, J. (2014). Why a propositional single-process model of associative learning deserves to be defended. *Dual Processes in Social Psychology*, 530–541.

De Houwer, J., & Beckers, T. (2003). Secondary task difficulty modulates forward blocking in human contingency learning. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *56*(4), 345–357. https://doi.org/10.1080/02724990244000296

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.

Deubel, H., & Schneider, W. X. (2003). Delayed Saccades, but Not Delayed Manual Aiming Movements, Require Visual Attention Shifts. *Annals of the New York Academy of Sciences*, *1004*(1), 289–296. https://doi.org/10.1196/annals.1303.026

Dodge, R. (1900). Visual perception during eye movement. *Psychological Review*, *7*(5), 454–465. https://doi.org/10.1037/h0067215

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455–470.

Escobar, M., Matute, H., & Miller, R. R. (2001). Cues trained apart compete for behavioral control in rats: Convergence with the associative interference literature. *Journal of Experimental Psychology: General*, *130*(1), 97.

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition*, *33*(6), 1036–1046.

Fernández, G., Castro, L. R., Schumacher, M., & Agamennoni, O. E. (2015). Diagnosis of mild Alzheimer disease through the analysis of eye movements during reading. *Journal of Integrative Neuroscience*, 1–13.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185. https://doi.org/10.1037/a0022100

Fernbach, P. M., Macris, D. M., & Sobel, D. M. (2012). Which one made it go? The emergence of diagnostic reasoning in preschoolers. *Cognitive Development*, *27*(1), 39–53.

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, *4*(1), 64–88.

Gallander Wintre, M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne*, *42*(3), 216–225. https://doi.org/10.1037/h0086893

Gopnik, A. (2009). Could David Hume Have Known about Buddhism?: Charles François Dolu, the Royal College of La Flèche, and the Global Jesuit Intellectual Network. *Hume Studies*, *35*(1/2), 5–28.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3.

Griffin, Z. M., & Oppenheimer, D. M. (2006). Speakers gaze at objects while preparing intentionally inaccurate labels for them. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 943.

Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, *42*, 279–297.

Higgins, E., Leinenger, M., & Rayner, K. (2014). Eye movements when viewing advertisements. *Frontiers in Psychology*, *5*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956003/

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795.

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: The role of uncertainty. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1658–1668.

Hogarth, L., Dickinson, A., & Duka, T. (2009). Detection versus sustained attention to drug cues have dissociable roles in mediating drug seeking behavior. *Experimental and Clinical Psychopharmacology*, *17*(1), 21.

Hogarth, L., Dickinson, A., & Duka, T. (2010a). Selective attention to conditioned stimuli in human discrimination learning: untangling the effects of outcome prediction, valence, arousal and uncertainty. *Attention and Associative Learning: From Brain to Behaviour*, 71–97.

Hogarth, L., Dickinson, A., & Duka, T. (2010b). The associative basis of cue-elicited drug taking in humans. *Psychopharmacology*, *208*(3), 337–351.

Hogarth, L., Dickinson, A., Hutton, S. B., Elbers, N., & Duka, T. (2006). Drug expectancy is necessary for stimulus control of human attention, instrumental drug-seeking behaviour and subjective pleasure. *Psychopharmacology*, *185*(4), 495–504.

Hume, D. (2003). *A treatise of human nature*. Courier Corporation.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, 105–134.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17. https://doi.org/10.1037/h0093874

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329.

Kamin, L. J. (1968). Attention-like" processes in classical conditioning. In *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–31).

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. *Punishment and Aversive Behavior*, 279–296.

Kant, I. (1848). Critique of Pure Reason , translated by Francis Haywood. *London: William Pickering*.

Klein, R. M. (1980). *Does oculomotor readiness mediate cognitive control of visual attention? In Nickerson (Ed), Attention and Performance (Vol. 8, pp. 259-276)*.

Klein, R. M., & Pontefract, A. (1994). Does Oculomotor Readiness Mediate Cognitive Control of Visual Attention? Revisited! *Attention and Performance XV: Conscious and Nonconscious Information Processing*, 333.

Kloos, H., & Sloutsky, V. M. (2013). Blocking a redundant cue: what does it say about preschoolers' causal competence? *Developmental Science*, *16*(5), 713–727.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. *Causal Learning: Psychology, Philosophy, and Computation*, 154–172.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(2), 220.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2014). Relative salience versus relative validity: Cue salience influences blocking in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*(1), 116.

Lindner, A., & Ilg, U. J. (2006). Suppression of optokinesis during smooth pursuit eye movements revisited: The role of extra-retinal information. *Vision Research*, *46*(6), 761–767.

Liversedge, S. P., White, S. J., Findlay, J. M., & Rayner, K. (2006). Binocular coordination of eye movements during reading. *Vision Research*, *46*(15), 2363–2374. https://doi.org/10.1016/j.visres.2006.01.013

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition*, *33*(8), 1388–1398.

Lucke, S., Lachnit, H., Koenig, S., & Uengoer, M. (2013). The informational value of contexts affects context-dependent learning. *Learning & Behavior*, *41*(3), 285–297.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276.

Matin, E. (1974). Saccadic suppression: a review and an analysis. *Psychological Bulletin*, *81*(12), 899.

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 182.

Matute, H., & Pineño, O. (1998). Stimulus competition in the absence of compound conditioning. *Animal Learning & Behavior*, *26*(1), 3–14. https://doi.org/10.3758/BF03199157

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183–198.

Mitchell, C. J., Griffiths, O., Seetoo, J., & Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(2), 191.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. 1998 http://www. usf. edu. *FreeAssociation.[PubMed]*.

Pavlov, I. P. (1927). Conditional reflexes: an investigation of the physiological activity of the cerebral cortex.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. Morgan Kaufmann Publishers, Los Altos.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction*, *1*, 211–219.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Quinlan, P. T. (2010). On the use of the term “attention.” *Attention and Associative Learning: From Brain to Behaviour*, 217–244.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, *45*(2), 245–260.

Shanks, D. R., & Lopez, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition*, *24*(4), 511–522.

Silverstein, S., Keane, B. P., Blake, R., Giersch, A., Green, M., & Kéri, S. (2015). Vision in Schizophrenia: Why it Matters. *Frontiers in Psychology*, *6*, 41.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103.

Spauschus, A., Marsden, J., Halliday, D. M., Rosenberg, J. R., & Brown, P. (1999). The origin of ocular microtremor in man. *Experimental Brain Research*, *126*(4), 556–562.

Sternberg, D. A., & McClelland, J. L. (2012). Two mechanisms of human contingency learning. *Psychological Science*, *23*(1), 59–68.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

Wagner, A. R., Logan, F., & Haberlandt, K. (1968). Stimulus selection in Animal discrimination learning. *Journal of Experimental Psychology*, *76*(2, Pt.1), 171–180. https://doi.org/10.1037/h0025414

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review*, *8*(3), 600–608.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the Information Given: Causal Models in Learning and Reasoning. *Current Directions in Psychological Science*, *15*(6), 307–311. https://doi.org/10.1111/j.1467-8721.2006.00458.x

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222.

Waldmann, M. R., & Walker, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior*, *33*(2), 211–229.

Yang, Q., Bucci, M. P., & Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, *43*(9), 2939–2949.

Appendix A.
Stimuli

Table 2

Causally-related word pairs for Experiments 1 and 2

Causally-related word pair		Forward str.	Backwards str.	2005 norms
Diet	Hunger	.01	.00	16
Spice	Flavor	.05	.02	-24
Virus	Epidemic	.00	nn	-24
Crime	Arrest	.00	.02	-20
Sadness	Crying	.05	.13	-20
Absence	Withdrawal	.01	nn	-12
Genes	Baldness	.00	.00	-10
Dairy	Diarrhea	.00	nn	-7
Spill	Stain	.00	.00	-7
Compliment	Blush	.00	.00	-4
Alcohol	Accident	.00	.00	-4
Panic	Escape	.00	.00	-3
Drought	Famine	.00	.02	-2
Joke	Amusement	.00	nn	-1
Bacteria	Infection	.01	.00	0
Attack	Defense	.05	.00	1
Birthrate	Population	nn	.00	1
Carcinogen	Tumor	nn	.00	2
Chromosome	Gender	.03	.00	2
Crush	Damage	.00	.01	3
Shock	Scream	.00	.00	3
Acid	Corrosion	.00	nn	5
Magnet	Attraction	.10	.04	6
Study	Pass	.00	.00	6
Lesion	Scar	nn	.00	7
Beat	Bruise	.00	.05	8
Pain	Aggression	.00	nn	10
Fracture	Cast	.00	.00	12
Training	Fitness	.00	.00	12
Illness	Treatment	.00	.04	12
Disease	Injection	.00	.01	13
Betrayal	Distrust	.00	.00	13

Sprain	Swell	.00	nn	18
Trash	Stink	.00	.00	20
Frowning	Wrinkles	.02	.00	24
Humidity	Sweat	.02	.00	27
Bang	Deafness	.00	nn	-11
Period	Cramps	.00	.07	4
Sunlight	Freckles	.00	.02	5
Frequency	Pitch	.00	.00	-3

Note. “Forward str.” and “Backwards str.” refer to the forward and backwards strength of association from the USF Word Norm Association list. “nn” indicates that the word pair has no norm in the USF database for that specific direction. “2005 norms” refers to the causal strength ratings given by participants in the causality norming study from Fenker, et. al (2005).

Table 3

Associated word pairs for Experiments 1 and 2

Associated word pair		**Forward str.**	**Backwards str.**
Graduation	Gown	.02	.00
Lime	Corona	.02	nn
Family	Sibling	.00	.03
Tuba	Saxophone	.03	.01
Propeller	Helicopter	nn	.03
Car	Plane	.01	.03
Insurance	Estimate	.00	.01
Test	Hypothesis	.00	.03
Power	Voltage	.01	.02
Chipmunks	Acorn	.00	.01
Antelope	Gazelle	.03	.03
Engine	Roar	.02	.00
Shrimp	Ocean	.03	.00
Caffeine	Mountain	.01	.00
Vessel	Vein	.01	.02
Cocktail	Fruits	.00	.01
Decency	Respect	.03	.00
Agency	Firm	.02	.00
Claw	Dogs	.01	.00
Ounce	Gallon	.02	.00
Envy	Admire	.03	.03

Bedroom	Furniture	.01	.00
Umbrella	Tote	.00	.01
Newspaper	Gossip	.00	.01
Session	Course	.01	.00
Computer	Apple	.01	.02
Patty	Hamburger	nn	.02
Sandwich	Tomatoes	.00	.00
Graph	Numbers	.02	.00
Glass	Window	.01	.02
Protestants	Baptist	.00	.01
Office	Employment	.00	.02
Atlas	Dictionary	.01	.00
Mother	Wife	.00	.03
Basketball	Teams	.00	.02
Terms	Meaning	.01	.00
Control	Volume	.00	.01
Acrobat	Athletes	.02	.00
Email	Attachment	nn	.00
Ambulance	Rush	.01	.00

Note. “Forward str.” and “Backwards str.” refer to the forward and backwards strength of association from the USF Word Norm Association list. “nn” indicates that the word pair has no norm for that specific direction. No causal strength was rated for associated word pairs in Fenker, et. al (2005).

Appendix B.

Task Instructions

After completing the calibration procedure for eye tracking, participants will see the following instructions in both Experiments 1 and 2:

> In this experiment, you will view a series of word pairs. Your task will be to determine whether the words are causally related. For each trial, you will first focus on the cross that will be displayed on the screen. Then, the first word will be displayed, followed shortly by the other. Press the C key if the first word causes or is caused by the second word. Press the N key if the words are not related.
>
> Since we will be recording how long it takes you to decide, please answer as quickly as you are able while getting as many answers correct as possible. To help you understand the task, you will complete 8 practice trials in which feedback will be provided before moving on to the experiment.

After the 8 practice trials, participants will see the following text to indicate that feedback will no longer be available after entering a response:

> For the following pairs of words, press:
>
> "C" if one of the words CAUSES the other
>
> OR
>
> "N" if the two words are simply associated
>
> -- Press SPACE to continue. --
>
> **No feedback will be given for these trials.**

Finally, after completing the task participants will see the following exit screen:

> THANK YOU FOR PARTICIPATING IN THIS STUDY!

Appendix C.

Analyses

Table 4

Experiment 1 Reaction Time (Log10 transformation)

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.01	1.78	0.19	0.06
A/C*Block	1	0.00	0.67	0.42	0.02
Error	29	0.00			
P/D	1	0.05	9.23	.005*	0.24
P/D*Block	1	0.00	0.62	0.44	0.02
Error	29	0.01			
A/C*P/D	1	0.05	11.90	.002*	0.29
A/C*P/D*Block	1	0.03	6.56	0.02*	0.18
Error	29	0.00			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at $p<.05$.

Table 5

Experiment 1 Accuracy

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.36	13.32	<.001*	0.31
A/C*Block	1	0.02	0.61	0.44	0.02
Error	30	0.03			
P/D	1	0.18	16.31	<.001*	0.35
P/D*Block	1	0.04	3.71	0.06	0.11
Error	30	0.01			
A/C*P/D	1	0.09	6.45	0.02*	0.18
A/C*P/D*Block	1	0.00	0.27	0.61	0.01
Error	30	0.01			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 6

Experiment 1 Alternations

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.41	5.19	0.03	0.15
A/C*Block	1	0.01	0.07	0.79	0.00
Error	30	0.08			
P/D	1	0.55	4.46	0.04*	0.13
P/D*Block	1	0.55	4.46	0.04*	0.13
Error	30	0.12			
A/C*P/D	1	0.70	5.05	0.03*	0.14
A/C*P/D*Block	1	0.01	0.08	0.78	0.00
Error	30	0.14			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 7

Experiment 2 Reaction Time (Log10 transformation)

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.02	1.06	0.32	0.06
A/C*Block	1	0.01	0.33	0.57	0.02
Error	16	0.02			
P/D	1	0.02	1.51	0.24	0.09
P/D*Block	1	0.05	3.32	0.09	0.17
Error	16	0.01			
A/C*P/D	1	0.00	0.65	0.43	0.04
A/C*P/D*Block	1	0.00	0.25	0.63	0.02
Error	16	0.01			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 8

Experiment 2 Accuracy

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.00	0.03	0.87	0.00
A/C*Block	1	0.01	0.16	0.70	0.01
Error	16	0.06			
P/D	1	0.08	11.39	.004*	0.42
P/D*Block	1	0.01	0.92	0.35	0.05
Error	16	0.01			
A/C*P/D	1	0.01	0.45	0.51	0.03
A/C*P/D*Block	1	0.02	1.52	0.24	0.09
Error	16	0.01			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 9

Experiment 2 Alternations

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.48	2.79	0.11	0.15
A/C*Block	1	0.08	0.46	0.51	0.03
Error	16	0.17			
P/D	1	0.28	3.40	0.08	0.18
P/D*Block	1	0.01	0.06	0.81	0.00
Error	16	0.08			
A/C*P/D	1	0.00	0.00	.998	0.00
A/C*P/D*Block	1	0.45	4.07	0.06	0.20
Error	16	0.11			

Note. Block = predictive first vs diagnostic first. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 10

Experiment 1 and 2 Log RT Comparison

Source	*df*	*MS*	*F*	*P*	*Partial η2*
A/C	1	0.03	3.09	0.09	0.06
A/C*EXP	1	0.00	0.27	0.60	0.01
Error	47	0.01			
P/D	1	0.07	7.04	0.01	0.13
P/D*EXP	1	0.00	0.05	0.83	0.00
Error	47	0.01			
A/C*P/D	1	0.04	7.12	0.01	0.13
A/C*P/D*EXP	1	0.01	1.53	0.22	0.03
Error	47	0.01			

Note. EXP = Experiment 1 vs. Experiment 2. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 11

Experiment 1 and 2 Accuracy Comparison

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.16	4.01	0.05*	0.08
A/C*EXP	1	0.11	2.75	0.10	0.06
Error	47	0.04			
P/D	1	0.23	21.92	<.001*	0.32
P/D*EXP	1	0.00	0.15	0.71	0.00
Error	47	0.01			
A/C*P/D	1	0.06	4.06	0.05*	0.08
A/C*P/D*EXP	1	0.01	0.99	0.33	0.02
Error	47	0.01			

Note. EXP = Experiment 1 vs. Experiment 2. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 12

Experiment 1 and 2 Average Alternations Comparison

Source	*df*	*MS*	*F*	*P*	*Partial* η^2
A/C	1	0.83	7.54	0.01*	0.14
A/C*EXP	1	0.04	0.34	0.56	0.01
Error	47	0.11			
P/D	1	0.00	0.00	0.97	0.00
P/D*EXP	1	0.69	6.20	0.02*	0.12
Error	47	0.11			
A/C*P/D	1	0.24	1.80	0.19	0.04
A/C*P/D*EXP	1	0.25	1.81	0.19	0.04
Error	47	0.14			

Note. EXP = Experiment 1 vs. Experiment 2. A/C = relationship type (associated vs. causal). P/D = causal order (predictive vs. diagnostic). Significant at p<.05.

Table 13

Experiment 1 and 2 Log RT post-hoc analysis

Trial Type	Exp. 1 Mean	Exp. 2 Mean	*t*	*df*	*p*
F. Associated	3.34	3.16	2.70	47	0.01*
B. Associated	3.34	3.17	2.50	47	0.017*
Predictive	3.32	3.17	2.32	47	0.025*
Diagnostic	3.40	3.22	2.51	47	0.016*

Appendix D.

Comparing Experiments 1 and 2

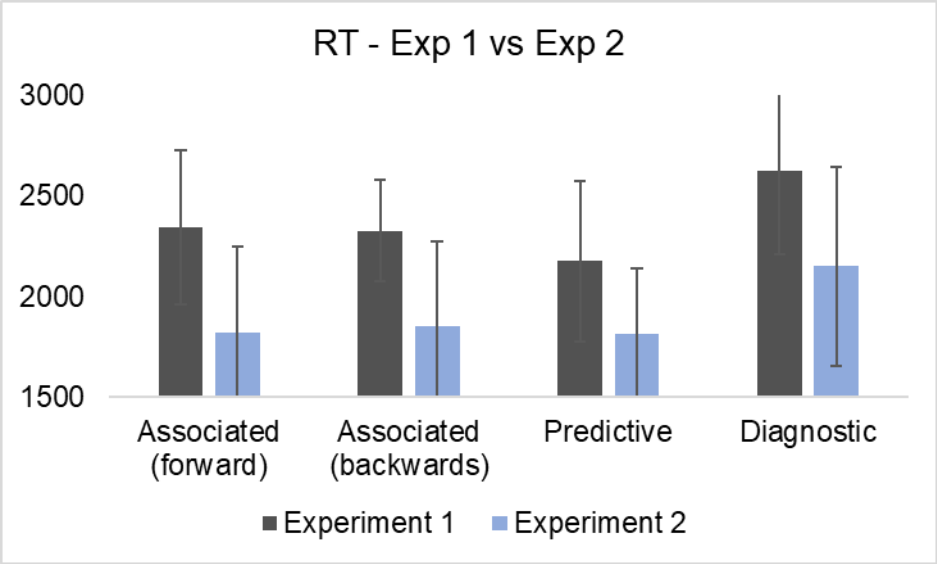

Figure 13. Reaction Time comparison between Experiments 1 and 2. Error Bars are *SEM*.

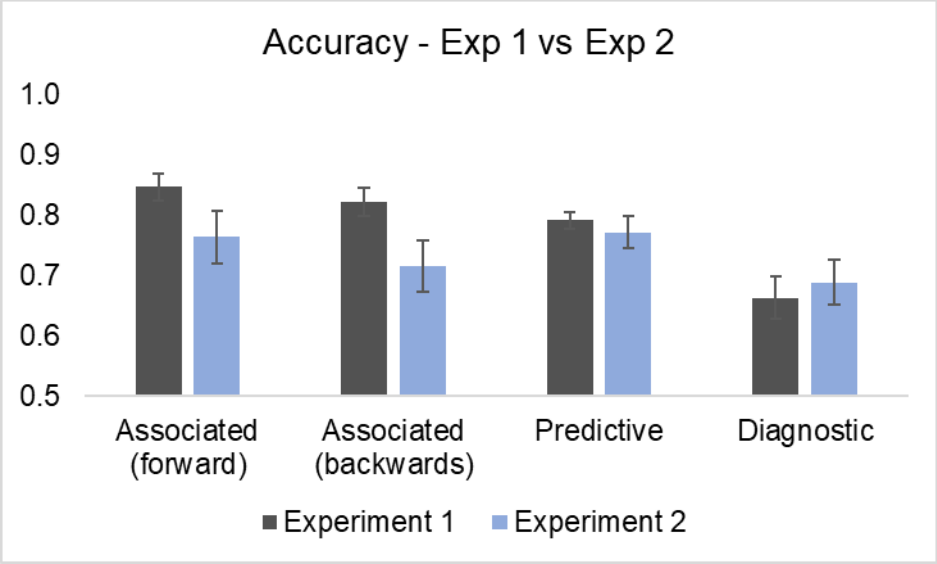

Figure 14. Accuracy comparison between Experiments 1 and 2. Error Bars are *SEM*.

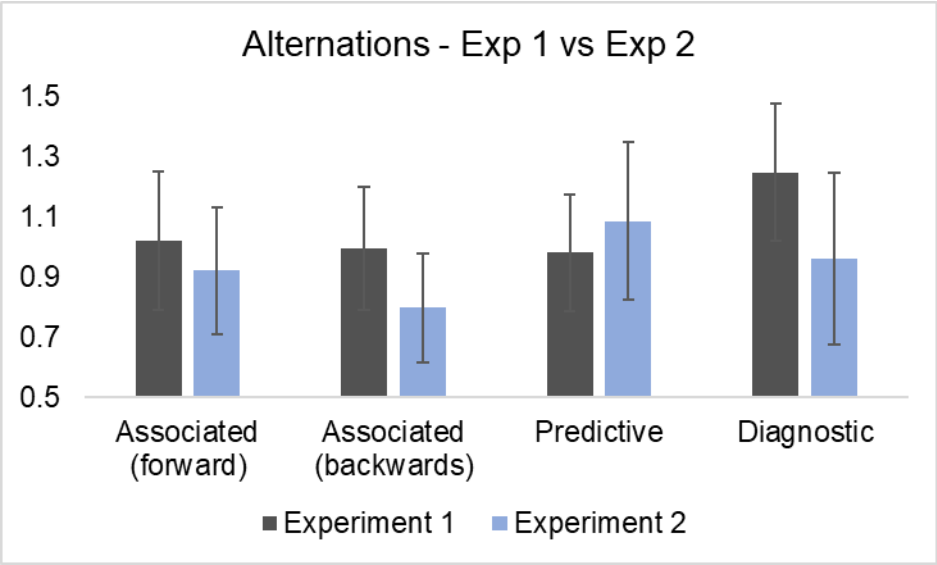

Figure 15. Comparison of average number of alternations between Experiments 1 and 2. Error Bars are *SEM*.