The Role of Granularity in Causal Learning

by

Kevin Wen Xin Soo

B.A. in Psychology, HELP University, 2009

M.S. in Cognitive and Decision Sciences, University College London, 2011

M.S. in Cognitive Psychology, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kevin Wen Xin Soo

It was defended on

March 26, 2019

and approved by

Melissa E. Libertus, Associate Professor, Psychology

Timothy J. Nokes-Malach, Associate Professor, Psychology, Learning Sciences & Policy

Christian D. Schunn, Professor, Psychology, Learning Sciences & Policy

James F. Woodward, Distinguished Professor, History & Philosophy of Science

Dissertation Director: Benjamin M. Rottman, Associate Professor, Psychology

The Role of Granularity in Causal Learning

Kevin Wen Xin Soo, PhD

University of Pittsburgh, 2019

Prior experiments on causal learning have typically investigated how people learn about the relationships between binary variables (e.g., patients either take or do not take a drug, and either exhibit or do not exhibit a particular symptom). Such experiments are often oversimplifications of real-world learning contexts, in which people have to learn about relationships between causes and effects of varying granularities (i.e. how many levels a variable has). In this dissertation, I explored how the granularities of a cause and effect influenced peoples' estimates of the strength of causal relationships. Four experiments were conducted in which participants learned about a cause-effect relationship by observing a cause and effect over multiple trials and making a judgment about the causal strength. On each trial, participants first viewed the state of the cause and predicted the state of the effect. Participants made stronger causal strength judgments when the effect was more coarse-grained, despite the objective causal strength being fixed (Experiment 1). The influence of the effect's granularity was due to participants perceiving the prediction task as subjectively easier when it involved a coarse-grained effect, and not due to feedback they received for their predictions (Experiment 2). These findings supported the newly proposed *feelings-of-success* heuristic; I proposed that participants made judgments of objective causal strength by substituting their subjective feelings of how successfully they made predictions of the effect. In support of this hypothesis, participants' judgments of how successful they were in the prediction task mediated the relationship between the granularity of the effect and their judgments of objective causal strength (Experiment 3). Finally, the influence of the effect's granularity was attenuated when

participants did not make explicit predictions, suggesting that the effect's granularity influenced causal strength judgments via the subjective feelings associated with the act of prediction (Experiment 4). Collectively, these studies show that while people are generally accurate when estimating causal strength, real-world factors like the granularity of variables can lead to biases in judgments.

Table of Contents

List of Tables

List of Figures

Preface

Everything good and true I have ever accomplished lies firmly on a causal chain of intertwining providence and blessing. I am eager to see where that chain goes. But first, I want to look back at the links that explain how I got here. These are the people I am indebted to.

My advisor, Benjamin Rottman. I believe it was the right decision for me to join him as his first ever graduate student. His brilliance, uncompromising ethics, and conscientiousness make him the best scientist I know. He taught me the virtues and habits I needed to be a rigorous researcher, and pushed me to produce better work than I believed I was capable of. The high bar he set for me was consistently matched by his generosity and willingness to go the extra mile in helping me get there. He knew when I needed a nudge or an encouraging word; at times it felt like he dragged me towards the finish line against my worst inclinations. Because of his mentorship, I am not only a better scientist, but also a better person. I will always be thankful for the honor of having worked alongside him, and grateful for the opportunities he has opened up for me. The future good work I do will have been caused in some way by him, but I take responsibility for any less-than-good side effects. As I take my first steps beyond academia, I have given up the privilege of being his colleague, but I am and always will remain his student.

My dissertation committee. That I now have a PhD is due in large part to the time and mental energy they gifted to shape both me and this dissertation. James Woodward and Christian Schunn offered questions and critiques that were both incisive and kind. Timothy Nokes-Malach and Melissa Libertus offered not only invaluable insights and feedback on my research, but also the professional guidance and wisdom I needed to navigate this stage of my career.

My peers and colleagues at the University of Pittsburgh and the Learning Research and Development Center, who formed a supportive and intellectually stimulating academic community. Emily Braham was my first friend at Pitt. Chris Olshefski enriched me with numerous lunchtime conversations. My lab-mates Ciara Willett and Zac Caddick brought much camaraderie and cheer to the end of my time in LRDC 720; I look forward to hearing of their future accomplishments in academia and beyond. Since my second year, Cory Derringer has been more than a colleague – he has been a constant friend (sometimes soccer teammate/fantasy football rival/travel companion); having him at the next desk made the hardest days a little better.

My church community. Joel Chan's friendship helped me grow a soul, back when I only possessed a mind. Matthew Koerber shepherded me through difficult seasons and released me with his blessing – few have cared for me better. Without my friends at Church of the Ascension, I would have been made a nihilist by the causal systems I inhabited; they wept with me, rejoiced with me, prayed with me, and hoped on my behalf when I could only despair. Father Jonathan Warren nurtured me with his friendship and spiritual direction; he taught me to see beyond the brass dome to a world charged with mystery. Grant and Christy Martsolf were radically hospitable and invited me to "come and see" the power of liturgical and sacramental worship; I am glad I did. The Janaszek-Hemphills welcomed me into their family and mentored me in kindness, thoughtfulness, and rock climbing; they showed me how to imagine and cultivate the good life – I expect to discover the fruits of their generosity for years to come. I will miss dinners with them.

Numerous friends and family in Pittsburgh (and beyond), who made it possible for me to put down roots halfway across the world. There are too many of them to name here. They welcomed me as one of their own; I am grateful for every single one of the hundreds of meals, conversations, and board game sessions we have shared over the years.

My mother, Angeline Lim, and brother, Timothy Soo. They taught me to consider others more important than myself, most powerfully by their own examples. They gave up a son and brother when I chose to pursue my professional ambitions halfway across the world, yet they have continued to support and love me in spite of the distance and obstacles between us. I am not convinced that my ambitions are worth the sacrifices I have asked of them, but we have to live with the choices we make. I strive to do so in a manner worthy of the grace they have shown me.

My father, Soo Ewe Jin. I wish he was here today. With each further step I take in life, I feel the void he has left behind. Growing up, I did not particularly enjoy it when he celebrated my accomplishments with his unique blend of fanfare and sentimentality, but I would give anything to see the look of pride on his face as I reach this milestone. He was a giant of a man, and a man worth making proud. He passed on many privileges and gifts that my life continues to be built upon, but none greater than the encouragement to see the world through my own eyes, even though his saw so much farther than mine.

My wife, Evelyn Yarzebinski. I am indebted to her above everyone else, and my accomplishment is hers also. Her care and patience for me as I wrote this dissertation was heroic. Difficult seasons such as these highlight the strength of her character in powerful ways, but more powerful still are the thousands of little, unseen, and mundane acts of love and faithfulness she performs out of habit. Every day, I find myself amazed at the kind and capable human being she is and continues to grow into, surprised at the transformative power of her belief in me, and humbled by the privilege I have of loving her. As we leave our life in Pittsburgh in this next season of life, I am grateful that I have her beside me. Only three years into our marriage, we have endured more sorrow and loss than I have known in all my years prior to meeting her. In that same time, we have also experienced more joy and contentment than I have ever known. It is evident to me that we are building a home that can weather hardship, and more importantly, a home that can receive, cultivate, and share good things. Maybe someday God will deem it right to bless us with the good things we most desire. In the meantime, I am looking forward to getting back to the good work of strengthening the foundations of our home, wherever that may be.

1.0 Introduction

In everyday life, we routinely learn and reason about the relationships between causes and their effects. Having knowledge about the relationships between such variables allows us to explain, predict, and exert control over our environments (Keil, 2006; Lombrozo, 2010; Sloman, 2005; Sloman & Lagnado, 2015; Waldmann & Hagmayer, 2014). Much research has explored how people learn about the strength of causal relationships from observational data (Cheng, 1997; Griffiths & Tenenbaum, 2005; for a review, see Holyoak & Cheng, 2011) and how people predict the state of a variable from knowledge of its causes or effects (Park & Sloman, 2013; Rehder, 2014; Rottman & Hastie, 2016; for a review, see Rottman & Hastie, 2013).

A prototypical causal learning task involves presenting participants with joint observations of a cause and an effect, and eliciting a judgment about the strength of the cause's influence on the effect. In most experiments, the cause and effect are typically represented as binary variables (e.g., multiple patients either take or do not take a drug, and they either exhibit or do not exhibit a particular symptom). While some variables do truly fall on a binary scale (e.g., a light may be switched either on or off), many variables can take on more than two values, and more importantly, we may perceive and represent such variables as having multiple levels when learning and reasoning about them (e.g., we can perceive gradations of temperature, as opposed to simply thinking of "hot" vs. "cold"). Thus, the variables used in experiments of causal learning and reasoning are sometimes oversimplifications of real-world variables, which can vary greatly in how fine-grained or coarse-grained they are.

The goal of the present research is to investigate causal learning and reasoning when the variables involved possess differing levels of *granularity* – i.e. the variables differ in the number of levels they can take on. The plan for the introduction is as follows. First, I provide an overview of prior research on causal learning that has ignored questions related to the granularity of the variables in order to motivate the present research. Second, I discuss the implications of causal learning with variables that are fine-grained vs. coarse-grained. Third, I outline several theories that make predictions for how the granularities of a cause and effect can influence causal learning. Fourth, I describe the ways causal strength judgments are measured in the present research. Finally, I provide an outline for the current experiments.

1.1 The granularity of variables in prior research

Most prior research on causal learning has focused on how people learn the strength of the relationship between a binary cause and binary effect (e.g., Barberia, Baetu, Sansa, & Baker, 2014; Buehner, Cheng, & Clifford, 2003; Liljeholm & Cheng, 2007; Mandel & Lehman, 1998; Shanks, Pearson, & Dickinson, 1989; White, 2003; for reviews, see Holyoak & Cheng, 2011; Perales & Shanks, 2007). This has been coupled with the development of normative models of causal induction that apply to binary causes and effects (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Jenkins & Ward, 1965; for a review, see Hattori & Oaksford, 2007).

Recognizing that real-world causal learning often involves combinations of both coarse- and fine-grained variables[1], more recent research has involved experiments involving binary causes and continuous effects (Chow, Don, Colagiuri, & Livesey, 2018; Derringer & Rottman, 2018; Obrecht, Chapman, & Gelman, 2007; Rottman, 2016; Saito, 2015; White, 2015), continuous causes and binary effects (Marsh & Ahn, 2009; Pacer & Griffiths, 2011), or continuous causes and continuous effects (Davis, Bramley, & Rehder, 2018; Soo & Rottman, 2016, 2018). The research cited here can be represented in a two-dimensional space with X as the granularity of the cause and Y the granularity of the effect, each of which range from binary to continuous.[2]

The increased focus on learning contexts involving more than just binary causes and effects leads to research with greater external validity, capturing the diverse granularities of variables found in real-world causal learning. However, no research to date has systematically investigated the implications of different levels of granularities of causes and effects on causal learning and

---

[1] The early and outsized focus on learning about relationships between binary variables was partly unique to the field of causal cognition. The same cognitive task – inducting a relationship between variables – has long been studied between continuous variables in domains such as function learning (Brehmer, 1971; DeLosh et al., 1997; Hogarth & Karelaia, 2011; Lucas, Griffiths, Williams, & Kalish, 2015; McDaniel & Busemeyer, 2005; Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017) and covariation detection (Crocker, 1981; Díez-Alegría, Vázquez, & Hernández-Lloreda, 2008; Erlick & Mills, 1967; Erlick, 1966; Hutchinson & Alba, 1997; Kareev, 1995; Lane, Anderson, & Kellam, 1985; Rensink, 2017; Vallée-Tourangeau, Baker, & Mercier, 1994; Vogel, Kutzner, Freytag, & Fiedler, 2014; Well, Boyce, Morris, Shinjo, & Chumbley, 1988).

[2] Although I refer to *continuous* variables, the variables used in such studies are not technically continuous because they do not have infinite possible levels. Rather, these variables can take on a very large number of possible levels such that participants in experiments perceive them as continuous. In the present research, I consider a variable to be continuous for practical purposes if it meets this criterion.

reasoning. This is a glaring lacuna; people often have to learn about the relationships between causes and effects that do not have similar granularities. For instance, people may have to learn about the relationship between a binary cause (e.g., either taking or not taking a painkiller pill) and a more fine-grained effect (e.g., varying levels of pain). In the present research, I investigated how the granularities of the cause and effect influence how people learn about the causal strength underlying a cause-effect relationship.

1.2 Continuous vs. discretized variables

As mentioned above, the binary variables used in causal learning experiments are sometimes oversimplifications of real-world variables. For example, while participants in experiments may be told a cover story involving patients who either take or do not take a drug, in the real world, drugs can be taken in dosages varying from zero milliliters ("no drug") to some maximum dosage in milliliters.

Some binary variables may truly have only two possible states in the real world, but binary variables (and variables with a discrete number of states more generally) may also be "discretized" from continuous variables. For example, the exact level of an enzyme in a patient's bloodstream is a continuous variable, but it may appear as a binary variable if it is measured by a test that is only sensitive to the presence vs. absence of the enzyme, or it may appear as a discrete variable with three levels, if the test is only sensitive enough to report low, average, or high levels of the enzyme. In other words, how fine- or coarse-grained a variable appears may be due to the sensitivity of its measurement.[3]

In the remainder of the present section, I describe a process by which continuous variables are discretized, and how this influences causal induction from a statistical perspective, potentially leading to issues for causal learning. Assume that a cause (*C*) and effect (*E*) are continuous variables related as described in the following equation:

$$E \sim f(C) + \varepsilon$$

The state of the effect is a function (*f*) of the state of the cause and noise (ε). For example, the level of a particular enzyme in peoples' bloodstreams (*E*) may be a linear function of the dosage of a drug (*C*) and noise. If 20 patients are administered different dosages of the drug, the causal relationship can be said to "generate" hypothetical data as in Figure 1. From all the observations, the correlation between the drug and enzyme levels (r = +.70) would provide an estimate of the strength of the causal relationship.

[3] The limits of measurement may be external to a learner (as in the example of tests with varying sensitivities), but may also be due to limits on a learner's perceptual system – i.e. learners may only be able to discriminate between levels up to a particular level of granularity. In the present research, I focus on the former case, in which variables appear to have particular granularities due to external limits of measurement.

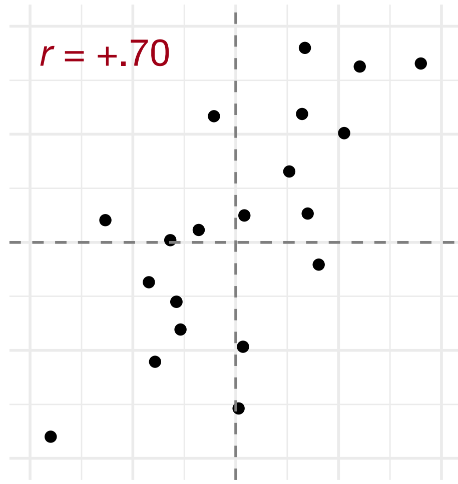

Figure 1. **Hypothetical data of 20 patients. A value of zero (one) indicates the minimum (maximum) of that variable. Dashed lines indicate median levels for both variables.**

However, consider what happens if the measurements of the drug and enzyme are only sensitive enough to detect if a substance is present or not; if the drug or enzyme exceeds a certain threshold (e.g., the median value of 0.50 out of 1.00) it is categorized as “present”, but is otherwise categorized as “absent” (see dashed lines in Figure 1). Such median splits are commonly used within the social sciences to categorize entities into groups based on their levels of a continuous predictor (Iacobucci, Posavac, Kardes, Schneider, & Popovich, 2015; MacCallum, Zhang, Preacher, & Rucker, 2002). This discretization scheme transforms C and E into binary variables. Other discretization schemes resulting in variables with more levels are also possible (e.g., Gelman & Park, 2009).

The practice of discretizing variables is sometimes necessary for practical reasons, but controversial from a statistical perspective (Cohen, 1983; Fitzsimons, 2008). Compared to continuous variables upon which they are based, discretized variables contain less information because they gloss over within-level variability (Rucker, McShane, & Preacher, 2015). Additionally, when estimating the strength of the relationships between variables, discretization can lead to false positives and false negatives (Maxwell & Delaney, 1993; McClelland, Lynch, Irwin, Spiller, & Fitzsimons, 2015). For example, if the data in Figure 1 were discretized based on the median cutoffs, the correlation between the discretized states of the cause and effect would be substantially weaker ($r = +.26$) than when it was computed using the continuous variables.

Although observations of continuous variables offer the best path to accurate causal induction from a statistical perspective, there are psychological reasons why human learners may show a preference for coarser-grained variables. Due to limitations in human perceptual and cognitive systems, learners may need to rely on discretized states of variables; early research from psychophysics has found that people are poor at identifying the absolute levels of continuous perceptual features (Garner & Hake, 1951; Helson, 1964; Restle & Merr, 1968; Sarris, 1967; Thurstone, 1927; also see Donkin, Rae, Heathcote, & Brown, 2015). Research has found that when presented with continuous variables, human learners spontaneously categorize observed continuous states into discretized states or categories (Fisher & Keil, 2018; Goldstone, 1994; Stewart, Brown, & Chater, 2002; Stewart & Chater, 2002). For example, when presented with tones of varying frequencies, people encode the relative levels between pairs of tones (i.e. which one was “higher” vs. “lower”) rather than the absolute levels (Stewart, Brown, & Chater, 2005). People use such discretized perceptual information to inform their higher-order judgments (e.g., Stewart, 2009; Stewart, Chater, & Brown, 2006), including causal judgments (Marsh & Ahn, 2009; Soo & Rottman, 2018).

In sum, when distinguishing between continuous vs. discretized variables, human learners appear to make better use of discretized information due to limitations in their perceptual and cognitive systems. However, from a statistical perspective, discretized variables inevitably lose some of the information captured by continuous variables. There appear to be competing reasons for why human learners might privilege variables with either finer or coarser granularity in causal learning.

1.3 How granularity influences causal learning

The present research is concerned with whether the granularities of causes and effects influence peoples' judgments of causal strength, when holding the objective causal strength constant. In the present section, I describe several theories that predict people might make either stronger or weaker judgments of the strength of causal relationships involving causes and effects of varying granularities. These theories are not mutually exclusive; it is possible that learners might display patterns of judgments consistent with multiple theories.

Even when the objective strength between the cause and effect is held constant (e.g., r = +.60), particular granularities in the cause and/or effect may influence learners to make stronger or weaker causal judgments than is warranted by the objective strength. Another way of saying this is that particular granularities in the variables may make a causal relationship appear "better" than it objectively is; causes that have a stronger influence on their effects can be viewed as being "better" because they have more utility for predicting and controlling their effects.

The following figure depicts the predicted causal judgments based on the theories presented here. Figure 2A displays the predicted causal judgments if a learner made judgments based only on the objective strength of the relationship; regardless of the granularities of the cause or effect, their judgments would be the same.

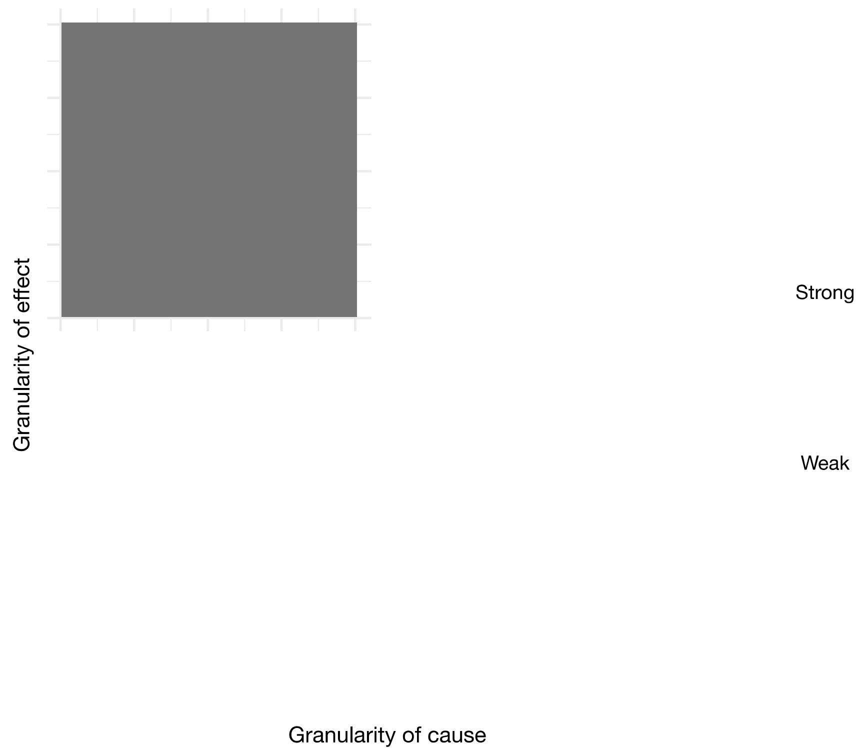

Figure 2. **Predicted causal judgments by various theories based on varying granularities of the cause and effect, when the objective strength of the causal relationship is held constant. Variables with lesser granularity (i.e. coarse-grained variables) have fewer levels.**

1.3.1 The granularity of the cause: The specificity criterion

One possibility is that the granularity of the cause could influence peoples' judgments of the strength of the relationship between the cause and the effect. There is a theory from philosophy that makes a somewhat analogous prediction, specifically, that causes containing the most detail (i.e. are more fine-grained) should always be judged as better causes (Franklin-Hall, 2016; Weslake, 2010; also see Griffiths, et al., 2015). According to this theory, coarse-grained variables are considered "epiphenomenal"; such variables are deemed to be mere aggregations or high-level descriptions of processes involving more fine-grained causes and effects. Thus, it is fine-grained variables, rather than coarse-grained ones, that exert causal powers over effects (Robb & Heil, 2018). I refer to this as the *specificity criterion* because it predicts a preference for causes with high specificity in the information they contain. Similar to the statistical perspective outlined in Section 1.2, this theory posits that finer-grained continuous variables are better than coarser-grained discretized variables for causal learning, except that it predicts that people are sensitive to the granularity of the cause, but not the effect.

This specificity criterion predicts an influence of the granularity of the cause on causal judgments. If people are sensitive to the specificity criterion, then they should make stronger judgments of causal strength when the cause is more fine-grained (see Figure 2B).

1.3.2 The granularity of the effect: The feelings-of-success heuristic

Another possibility is that the granularity of the effect could influence peoples' judgments of the strength of the relationship between the cause and the effect. Although I am unaware of any philosophical theories that make this prediction, there are psychological reasons for making this prediction, specifically, that people will judge the cause-effect relation to be stronger when the effect is more coarse-grained. This hypothesis has to do with the fact that (1) people spontaneously make predictions based on related information that they observe, and (2) causes occur before their effects; upon observing a cause, people spontaneously make predictions about related effects (Buehner, 2012; Fernbach, Darlow, & Sloman, 2011; Greville & Buehner, 2010; Lagnado & Sloman, 2006; Shanks et al., 1989; Sloman & Lagnado, 2015, 2005).

The tendency to make predictions as described in (1) applies to many domains of cognition. When given information about some variable, people automatically and intuitively make inferences about related quantities. These inferences include making predictions about behaviors and membership in social groups (Kahneman & Tversky, 1973; McCauley, Stitt, & Segal, 1980; Nisbett & Borgida, 1975; Sarbin, 1944; Tversky & Kahneman, 1983), predictions of numerical quantities from cues (Birnbaum, 1976), predictions of future words when reading (Federmeier, 2007), and predictions of the consequences of motor actions (Knoblich & Flach, 2001; Sebanz & Knoblich, 2009; Wolpert, Doya, & Kawato, 2003). These constitute a more active form of learning about the world (as opposed to passive observation); people make predictions based on information they possess, and either confirm or update their beliefs based on the accuracy of their predictions (see Danks, 2003; Miller, Barnet, & Grahame, 1995; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972).

The tendency to make predictions is particularly relevant when learning about causal relationships in the real world, because as described in (2), people commonly observe causes before their effects. After experiencing or observing a cause, people form expectations about the subsequent appearance of its associated effect (Buehner, 2012; Hagmayer & Waldmann, 2002; Lagnado & Sloman, 2004).

Within a learning context in which people have the opportunity to observe a cause and make subsequent predictions of an effect, it is possible that the granularity of the effect influences causal judgments. Learners may find it more difficult to make precise predictions of more fine-grained effects because they are less likely to predict the effect's exact true state. In contrast, making predictions of more coarse-grained effects may be perceived as easier because learners are likelier to predict the effect's exact true state, even if only by chance.

In light of this, I propose a novel hypothesis: learners judge the strength of causal relationships based on their feelings of perceived ease (or difficulty) in making accurate predictions of the effect during learning, which I refer to as subjective feelings-of-success.

Specifically, I propose that learners will experience greater subjective feelings-of-success when making predictions of more coarse-grained effects, and these feelings will then be substituted for judgments of the objective causal strength. I refer to this substitution as the *feelings-of-success heuristic*. The use of this heuristic should result in stronger judgments when the effect is more coarse-grained (see Figure 2C). Another way of describing the substitution implied by the heuristic is to say that subjective feelings-of-success is a mechanism or mediator by which the granularity of the effect influences judgments of objective causal strength.

The feelings-of-success heuristic is analogous to instances in which people make judgments of inaccessible quantities by "substituting" more accessible affective and subjective quantities. For example, people use feelings of fluency as a substitute for their judgments of how well they have learned some information (Hertwig, Herzog, & Schooler, 2008; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat, 2008; Koriat & Ma'ayan, 2005; Oppenheimer, 2008). As another example, people judge cities they recognize as being larger than cities they do not recognize; this substitution is reasonable if one assumes that larger cities tend to be more recognizable or salient in memory (Gigerenzer, 2008; Gigerenzer & Goldstein, 2011; Goldstein & Gigerenzer, 2002; Kahneman & Tversky, 1973; Pachur & Hertwig, 2006).

Subjective feelings-of-success can be viewed as a rational cue to objective causal strength in most situations. When a causal relationship is very strong (e.g., between the cause and effect, $r = +1$ or -1), a learner will presumably be able to make accurate and successful predictions. If the causal strength is weak (e.g., r close to zero), a learner will presumably have less accuracy and success in making predictions. Thus, prediction accuracy (and consequently, prediction success) should be highly correlated with the objective causal strength of a relationship in most situations.

In the present research, the objective causal strength is held constant, and the question of interest is whether manipulating the granularity of the effect might artificially influence learners' subjective feelings-of-success, and thus their judgments of objective strength.

1.3.3 The granularity of the cause and effect: The proportionality criterion

In principle, the granularities of the cause and effect could interact in various ways, in turn influencing peoples' judgments about causal strength. I focus on one particular theory-driven hypothesis, that people might make stronger judgments of causal strength if the granularities of the cause and effect are the same.

Within the field of philosophy, features of variables akin to the granularity of variables can be taken as criteria for evaluating causal claims. Woodward (2010, 2018a, 2018b; also see McGrath, 1998; Yablo, 1992) refers to this as the *proportionality criterion*. The logic of it is as follows. Causal claims about some system (e.g., that *C* is a strong cause of *E*) can be framed at different levels of description, each involving variables that are more or less detailed. According to Woodward, causal claims should be judged more favorably if the cause contains all necessary information relevant to explaining the effect, but omits all unnecessary information. In other words, Woodward is claiming that causes that are proportional to their effects (i.e. similar in its level of description) should be endorsed as good causes for predicting and controlling the effect.

The proportionality criterion is a general principle that can be applied broadly to assess what constitutes an appropriate causal inference. For example, a colleague's behavior (e.g., writing a curt email, or being late to a meeting) can be explained at (1) a fine-grained physiological level, as being due to particular patterns of neuronal activity, or (2) a coarse-grained psychological level, by appealing to cognitive concepts like intentions and beliefs. Based on the proportionality criterion, a cause from level (2) would be judged as a better cause, as it provides a causal explanation for the effect that is situated at a level of description that is more proportional to its effect.

By analogy, the proportionality criterion prescribes that causes will be judged stronger if they are similar to the effect in granularity. In other words, if people are sensitive to the proportionality criterion, then causes that are proportional to their effects should receive stronger causal judgments. The plot for the proportionality criterion (Figure 2D) displays stronger predicted causal judgments on the diagonal representing combinations of causes and effects that are proportional in granularity.

To understand why learners might make stronger causal strength judgments when the cause and effect are proportional, consider a scenario in which the cause and effect both have five levels. In this scenario, learners could naively map each level of the cause to the corresponding level of the effect and interpret that one-to-one mapping as the definition of a positive causal relationship (alternatively, they could map the highest level of the cause to the lowest level of the effect, the second highest level of the cause to the second lowest level of the effect, and so forth, and interpret that as a negative causal relationship). This could make it easier to see how the cause and effect are related, and consequently lead to stronger causal strength judgments. In contrast, if the cause had two levels and the effect had five levels, it would be harder to map the states of the cause to states of the effect, making it more difficult to see how the cause and effect are related, leading to weaker causal strength judgments.

1.4 Measuring causal strength

In the present research, I investigate the role of granularity using a common learning paradigm in which people learn about the strength of the relationship between a cause and effect by observing data of those variables presented sequentially (e.g., Luhmann & Ahn, 2007, 2011; Soo & Rottman, 2018; Speekenbrink & Shanks, 2010). In some prior experiments using this paradigm, learners are asked to make explicit predictions of the effect on each trial based on the observed state of the cause (e.g., Derringer & Rottman, 2018; Spellman, 1996). In others experiments, learners do not make explicit predictions, although they may make implicit predictions, provided the state of the cause is observed prior to the effect on each trial.

After observing all the data for a particular cause-effect pair, learners make judgments about the strength of the causal relationship. In the following sections, I describe the three measures I used to investigate learners' assessments of the causal relationship.

1.4.1 Measures of the strength of the relationship between the cause and effect

The target of interest in typical causal learning tasks is the learner's estimate of the objective strength of the relationship between the cause and effect; the correlation between the observed states of the cause and effect can be considered the normative benchmark for this estimate.

I decided to use two different measures of causal strength, reflecting two different views of the purpose of regression. One purpose of regression is to assess the strength of a relationship between two variables – to measure the "variance explained" and whether the relation is statistically significant. Another purpose of regression is to be able to predict the dependent variable given the independent variable(s). That is, once the regression model has been fitted to existing data, one can use the fitted parameters to make predictions about new cases. These two purposes of regression are mathematically equivalent – the stronger the relationship, the more accurate predictions will be, and vice versa. However, they are often spoken of as distinct functions of regression and computational models more generally (see Shmueli, 2010). Given that there are two ways to think about regression, it was possible that when asking learners for their assessments of causal strength, that one of these framings might be easier to understand than the other, or that they might be interpreted differently.

The first measure I used is a traditional measure of *causal strength*, which is commonly used in causal learning experiments (e.g. Cheng, 1997; Cheng & Novick, 1992; Soo & Rottman, 2018). The *causal strength* measure typically elicits learners' estimates of causal strength using a Likert scale with the negative end of the scale indicating a very strong negative causal relationship, the positive end indicating a very strong positive relationship, and the middle of the scale indicating there is no relationship.

The second measure I used to assess learners' estimates of causal strength is the extent to which learners believe that the cause actually predicts the effect, which I refer to as *predictiveness*. This measure requires learners to assess how well the actual states of the cause seemed to predict the actual states of the effect that were observed in the data, ignoring the predictions they may have made of the effect during learning (the following section describes a separate measure of how accurate learners' predictions of the effect were). The *predictiveness* measure elicits learners' estimates of the causal relationship using a Likert scale with the low end of the scale indicating the cause is "not predictive" of the effect, and the high end of the scale indicating the cause is "very predictive" of the effect. Unlike the *causal strength* measure, the *predictiveness* measure only captures the strength of the association while ignoring the polarity of the relationship (positive vs. negative).

1.4.2 A measure of subjective feelings-of-success

As mentioned above, learners observing sequential data of a cause and effect will often make predictions – either explicitly or implicitly – of the effect based on the state of the cause observed on each trial. In addition to the measures of objective causal strength described above, another relevant quantity when learners are making predictions is the subjective feelings-of-success they experience based on the accuracy of their predictions. Whenever learners make predictions about the state of the effect, they subsequently see the actual state of the effect, which provides them with feedback about their accuracy.

I introduced the *prediction success* measure to capture learners' subjective feelings-of-success from making predictions of the effect across all trials. This measure elicited learners' subjective feelings-of-success using a Likert scale with the low end of the scale indicating they were "not successful" in predicting the effect, and the high end of the scale indicating they were "very successful" in predicting the effect. This measure is akin to measurements of task performance (see the NASA Task Load Index; Hart & Staveland, 1988).

In contrast to the measures above that target the objective strength of the relationship (and are based on the correlation between the actual states of the cause and effect), the measure of *prediction success* is based on learners' perceptions of how accurately they predicted the state of the effect from the cause. One reason for measuring subjective feelings-of-success was because it might be related to learners' judgments of objective causal strength; in particular, the hypothesis that learners use the feelings-of-success heuristic (Section 1.3.2) predicts that subjective feelings-of-success might be a mediator of objective causal strength judgments. Measuring *prediction success* allowed the testing of this hypothesis.

1.4.3 Hypotheses about the influence of granularity on these three measures

The theories outlined in Section 1.3 predict that learners' causal strength judgments will be influenced by the granularities of the cause and/or effect, when holding the objective causal strength constant.

First, if learners are sensitive to the specificity criterion, they will make stronger judgments of objective causal strength (using the *causal strength* and *predictiveness* measures) when the cause is more fine-grained. It is also possible that fine-grained causes will lead to learners feeling that they can more accurately predict the effect, leading to greater subjective feelings-of-success, and consequently, judgments of *prediction success*.

Second, if learners use the feelings-of-success heuristic, they will make stronger causal strength judgments when the effect is more coarse-grained. This prediction applies primarily to learners' judgments of *prediction success*; learners should experience greater subjective feelings-of-success when predicting more coarse-grained effects because they are more likely to predict the effect's exact true state. However, I additionally predict that if learners use the feelings-of-success heuristic, they will make judgments of objective causal strength based on their subjective feelings-of-success. Thus, this hypothesis predicts that learners' judgments using all measures will be greater when the effect is more coarse-grained.

Finally, if learners are sensitive to the proportionality criterion, they will make stronger judgments of objective causal strength when the cause and effect possess the same granularities. In addition, it is possible that learners will also make greater judgments of *prediction success* because making predictions between proportional causes and effects may lead to greater subjective feelings-of-success because learners can easily map levels of the cause to levels of the effect.

1.5 Outline of research

The goal of the present research was to investigate if and how the granularity of the cause and effect influence judgments about the strength of the relationship between a cause and effect. Across four experiments, participants experienced data of causes and effects in which the strength of the causal relationships was fixed. Participants learned the strength of the relationships while making predictions of the state of the effect after viewing the actual state of the cause.

In Experiment 1, I manipulated the granularities of the cause and effect to determine how they influenced participants' causal judgments. In Experiment 2, I investigated whether the influence of the granularities of the cause and effect were present when participants either did or did not receive an accuracy bonus for their predictions. In Experiment 3, I tested if participants' subjective feelings-of-success mediated the relationship between the granularity of the variables and participants' judgments of the objective strength of the causal relationship; in other words, I tested subjective feelings-of-success as a mechanism. In Experiment 4, I further tested the mechanism by investigating how the influence of the granularity of the variables was moderated by whether or not participants engaged in prediction.

2.0 Experiment 1: The influence of granularity on causal learning

In Experiment 1, I investigated the role of a cause and effect's granularity on causal learning. Participants learned about causal relationships by observing the states of a cause, and predicting the resulting states of an effect. After experiencing the data for a particular cause-effect relationship, participants made judgments of causal strength based on their observations. In each dataset, the cause and effect had the same objective correlation, but the number of possible levels of each variable was manipulated across datasets. The goal of this experiment was to determine if the granularity of the cause and/or the effect influenced peoples' causal judgments.

Different patterns of results would imply different criteria that people may be sensitive to when making causal judgments. If participants' judgments are stronger when the cause is more fine-grained, participants' judgments would be sensitive to the specificity criterion, i.e. that causes with more levels are perceived to be more informative and specific. If participants' judgments are stronger when the effect is more coarse-grained (i.e. has fewer levels), this would be evidence that participants were making judgments based on their subjective feeling of how accurately they are using the cause to predict the effect, which I call the feelings-of-success heuristic. Finally, if participants' judgments are stronger when the cause and effect have the same number of levels, participants' judgments would be sensitive to the proportionality criterion. The various predictions are not mutually exclusive; it is possible that participants will be sensitive to multiple criteria.

2.1 Method

2.1.1 Participants

120 participants were recruited using Amazon Mechanical Turk (Mturk). Data from one participant was not recorded due to a programming error. All participants lived within the United States, had at least 100 approved assignments on Mturk, and had an assignment approval rate of at least 95%. Research on online labor markets has found these criteria help ensure high quality data in behavioral experiments (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014). Each participant was paid \$2.00, and an additional accuracy bonus (*M* = \$1.07, *SD* = \$0.10) based on their performance in the trial-by-trial prediction task. The experiment lasted approximately 12-15 minutes.

2.1.2 Design

The experiment utilized a 3 (number of levels in the cause: 2 vs. 5 vs. 13; within-subjects) × 3 (number of levels in the effect: 2 vs. 5 vs. 13; within-subjects) × 3 (measure: *causal strength* vs. *predictiveness* vs. *prediction success*; between-subjects) design. The number of levels in the cause and effect ($Levels_C$ and $Levels_E$, respectively) were manipulated by varying the number of levels that each variable could take on in a particular dataset observed by participants, resulting in nine scenarios per participant. The measure denotes the particular response scale participants used to make their judgments at the end of each scenario. Forty participants were assigned to each between-subjects condition.

2.1.3 Datasets

In each scenario, participants viewed a dataset consisting of 20 joint observations of a cause and effect. Each variable could take on values ranging from one to the number of levels of that variable (Levels_C and Levels_E). The correlation (Pearson's r) between the cause and effect in all datasets was fixed to be within a small range, so that $r = .60 \pm .01$.

I created datasets via the following processes. Datasets in which the cause and effect were both binary (i.e. $\text{Levels}_C = \text{Levels}_E = 2$) always had the same 20 joint states (Table 1) so that the association between the variables was fixed at $r = .60$.[4] Two hundred unique datasets were created for this condition by randomizing the order of observations for this dataset.

Table 1. ***Frequencies of Joint States in Datasets with Binary Variables***

		State of effect, E	
		$E = 1$	$E = 2$
State of cause, C	$C = 1$	8	2
	$C = 2$	2	8

Note. The absolute states of "1" and "2" are arbitrary, but are meant to imply the smallest vs. largest possible state given a particular variable's scale.

In all other conditions, the following process was used to generate the datasets. Using the MASS package in R (Venables & Ripley, 2002), I generated datasets consisting of 20 observations of two continuous variables, with $r = .60$. I then created multiple versions of each dataset by

[4] Since the variables are binary, r is equivalent to the phi coefficient ($\varphi = .60$).

discretizing the values of each variable into two, five, or 13 levels.[5] After discretization, datasets in which the correlation between variables was $r = .60 \pm .01$ were retained to be used as stimuli.[6] Datasets were generated until 200 unique datasets in each condition were obtained.

I created a copy of each dataset in which the states of the cause were reverse-coded along the midpoint of the scale. This resulted in datasets for which the correlation between variables was negative ($r = -.60 \pm .01$). This was done so that participants were not always viewing data with the same moderately strong positive correlation, which might have been predictable and boring. Using datasets with varying correlations would also help distract from the main manipulation of the granularities of the cause and effect.

---

[5] A two-level (binary) variable is the most coarse-grained variable possible. I considered 13 levels to be fine-grained enough that for practical purposes it was close enough to a continuous variable. A five-level variable was somewhere between these extremes, although closer to a two-level variable. This was chosen because people were likely more sensitive to granularity at the low end of the scale; the influence of granularity necessarily tapers off at the high end of the scale as the granularity increases to the point that people perceive the variable as continuous. The chosen granularities were meant to capture important points on the scale. Furthermore, the possible granularities of the cause and effect were not multiples or factors of each other (two, five, and 13 levels) to ensure that the states of the cause and effect could not be easily mapped onto each other when the granularities differed.

[6] There was a second criterion for a dataset to be retained as stimuli. The relationship between both variables had to be linear, rather than a higher-order polynomial relationship. Appendix A contains details of how this was assessed in all datasets.

2.1.4 Procedure and cover story

At the start of the experiment, participants were given instructions on navigating the experiment website and completed a guided tutorial of one scenario before viewing the actual scenarios.

Participants were told they were researchers studying how drugs influenced the levels of enzymes in peoples' bloodstreams. Each scenario involved learning about the relationship between one particular drug (the cause) and one particular enzyme (the effect), and subsequent scenarios involved different drug-enzyme pairs. Participants were told that if the drug had a positive (negative) causal influence on the enzyme, being administered a low dosage of the drug would result in low (high) levels of the enzyme, an average dosage would result in average levels of the enzyme, and a high dosage would result in high (low) levels of the enzyme. Participants were also told it was possible a drug had no causal influence on the enzyme, in which case particular dosages would not reliably lead to particular levels of the enzyme. These descriptions of the relationship were meant to imply a linear relation between the cause and effect.

To learn the relationship between a drug-enzyme pair, participants observed data from a clinical trial in which 20 volunteers were administered different dosages of the drug, and then had the level of the enzyme measured in their blood. Each volunteer's drug dosage and corresponding enzyme level constituted a single observation in the dataset.

The states of the drug and enzyme were represented using two vertical gauges with varying numbers of regions depending on Levels_C and Levels_E (see Figure 3); the lowest region on the gauge indicates the lowest state. The cause (effect) was represented by the gauge on the left (right). The state of a variable on a given trial was indicated by filling the corresponding region light blue.

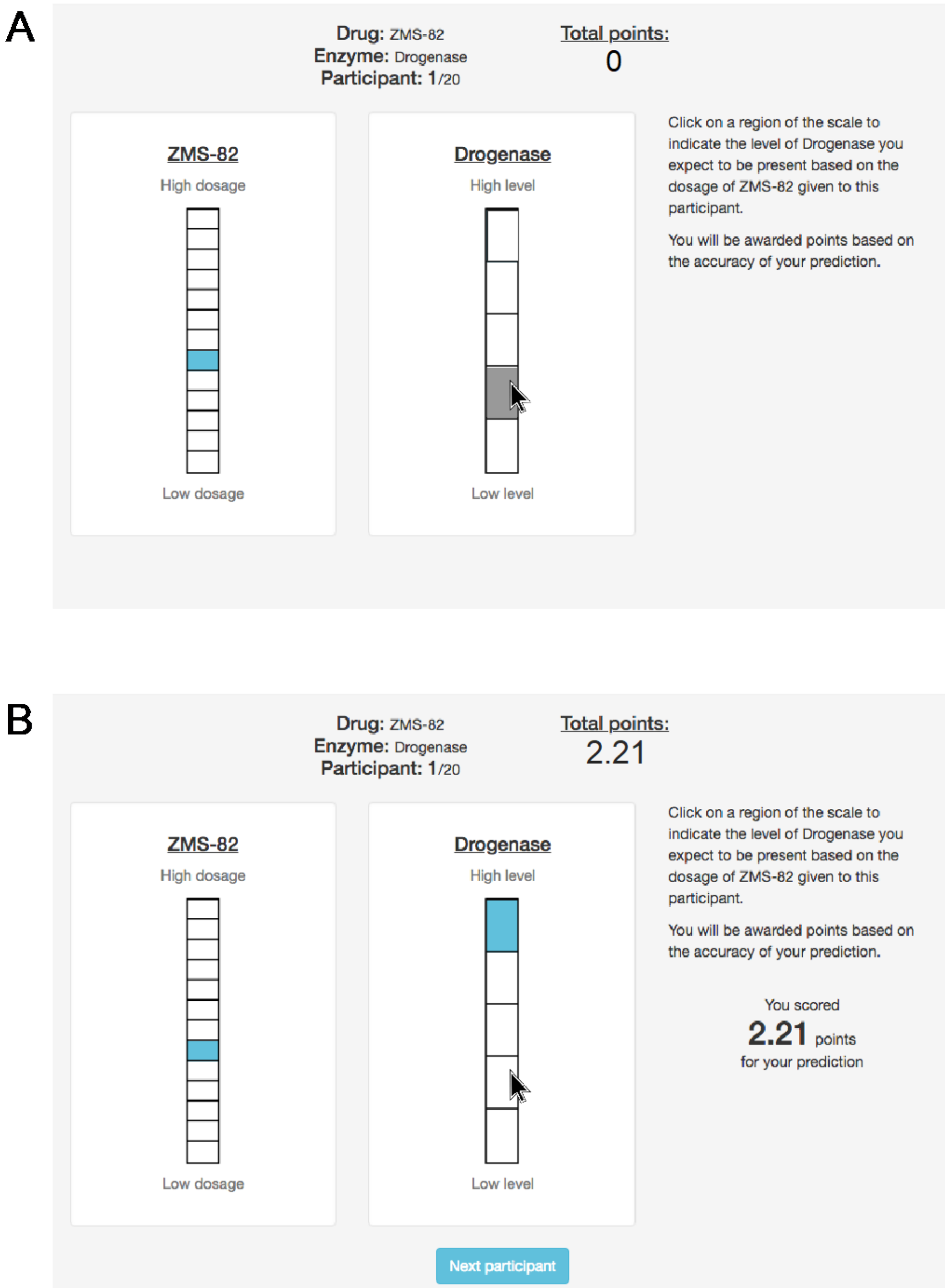

Figure 3. **Presentation of stimuli in Experiment 1. In this condition, Levels$_C$ = 13 and Levels$_E$ = 5. (A) After being shown the state of the cause (indicated by the light blue region in the gauge representing the cause), the participant predicts the state of the effect by clicking on a region in the gauge representing the effect, which flashes gray. (B) The correct state of the effect is revealed, and the points won for the participant's prediction (2.21) is displayed and added to the "Total points" above, which keeps track of all points won during the experiment. The newly visible button below the stimuli allows participants to advance to the next trial.**

Participants were told that the granularity of a particular variable reflected the sensitivity of the test used to measure that variable – i.e. a sensitive test meant that fine-grained distinctions could be made between the states of that variable, corresponding to gauges with more levels.

During each scenario, participants learned about the relationship between the cause and effect by engaging in a trial-by-trial prediction task. On each trial, after clicking on a button to display the observation for that particular trial, participants were first shown the state of the cause; a particular region of the gauge on the left was shaded light blue. Participants then had to predict the corresponding state of the effect by clicking on a particular region of the gauge on the right, which flashed briefly. After making their prediction, the actual state of the effect was shown; the region of the gauge corresponding to the correct state of the effect was shaded light blue. Participants were then awarded a bonus based on the accuracy of their prediction (Section 2.1.5). The bonus for that trial was flashed briefly on the right-hand side of the stimuli and then added to the experiment-long bonus total above the stimuli (see Figure 3).

After experiencing all 20 observations, participants made a judgment concerning the entire dataset using one of the measures (described in Section 2.1.6). Participants then progressed to the following scenario, until they experienced all nine scenarios.

2.1.5 Accuracy bonus

After predicting the effect, in addition to being shown the true level of the effect, participants were also given a bonus based on the accuracy of their prediction (Figure 3B). On each trial, participants were awarded between 0.00 and 10.00 points. At the end of the experiment, participants were paid one cent for every 10 points (rounded down), in addition to their base payment. Participants were told of this bonus payment rate prior to starting the experiment.

Accuracy bonuses were awarded for two reasons. First, providing bonuses would presumably motivate participants to pay attention, learn the relation between the cause and the effect, and provide accurate predictions. Second, I was concerned that participants might adopt a misconceived notion of prediction accuracy. In particular, imagine if a participant only counted a prediction as being accurate if they predicted the exactly correct region, and they counted a close prediction as wrong. Such a notion of prediction accuracy would mean that when the effect only has two levels, participants would often get the prediction exactly correct, but when it has 13 levels, they would rarely predict the exact effect. I wanted to use a bonus scheme that encouraged subjects to internally measure *prediction success* in a way that is consistent with how the stimuli were created – linear regression – to ensure that the conditions are in fact comparable. Below, I explain how the bonuses were determined.

First, I standardized the cause and effect variables in all datasets across all conditions to put them on the same scale. Second, within each condition, I fitted regression models to each of the 200 dataset. Third, I used the regression parameters to predict the effect from the cause. Fourth, because these predictions do not fall exactly onto one of the levels of the effect variable, I rounded these predictions to the closest level of the effect. Fifth, for each prediction, I computed the squared error between the rounded prediction and the actual effect. Because this is on the standardized scale, I call this error the "standardized squared error" (SE_{Std}). Sixth, I also recorded the raw error of that prediction if the standardized predictions were transposed back to their original scale, and took the average SE_{Std} that corresponds with 0, 1, 2, etc., levels of error on the raw scale (2, 5, or 13). Table 2 displays how raw prediction errors are associated with different levels of SE_{Std} for all nine conditions. Now, when the effect only has two levels, a raw error of one is associated with a large SE_{Std} but when the effect has five or 13 levels, a raw error of one is associated with a much smaller SE_{Std}.

Table 2. ***Accuracy Bonus Scoring Scheme***

	Levels$_C$ = 2		Levels$_C$ = 5		Levels$_C$ = 13	
Raw error	Mean SE_{Std}	Points	Mean SE_{Std}	Points	Mean SE_{Std}	Points
			Levels$_E$ = 2			
0	0.15	8.50	0.27	8.50	0.27	8.50
1	2.43	0.00	1.8	0.00	1.79	0.00
			Levels$_E$ = 5			
0	0.04	9.58	0.04	9.58	0.04	9.58
1	0.51	6.61	0.51	6.64	0.50	6.67
2	1.69	3.72	1.74	3.65	1.74	3.65
3	3.42	2.26	3.49	2.23	3.53	2.21
4	–	0.00	–	0.00	5.47	1.55
			Levels$_E$ = 13			
0	0.01	9.94	0.01	9.94	0.01	9.94
1	0.07	9.31	0.08	9.29	0.08	9.29
2	0.28	7.80	0.29	7.78	0.28	7.81
3	0.63	6.14	0.62	6.17	0.62	6.18
4	1.09	4.79	1.10	4.76	1.10	4.77
5	1.71	3.69	1.69	3.72	1.69	3.72
6	2.38	2.96	2.42	2.92	2.41	2.93
7	3.19	2.39	3.23	2.36	3.26	2.35
8	4.35	1.87	3.75	2.11	3.99	2.00
9	4.82	1.72	5.02	1.66	5.00	1.67
10	–	0.00	–	0.00	–	0.00
11	–	0.00	–	0.00	–	0.00
12	–	0.00	–	0.00	–	0.00

Note. SE_{Std} from regressions are computed based on averaging predictions from all 200 datasets that correspond to different levels of raw prediction errors. In some conditions, some large raw errors are awarded zero points because the regressions never made predictions with that amount of SE_{Std} for any datasets.

The last step was to determine how to convert the SE_{Std} into a bonus. I decided to award bonuses that could range from 0 to 10.00 points, based on the following equation:

$$Points_t = \frac{10.00}{(SE_{Std} + 1)}$$

I wanted to verify that the bonuses earned would be roughly equivalent across all of the conditions assuming that participants performed equally well across conditions, so I computed the sum of the bonus that would be earned using linear regression to predict the outcomes. Most of the nine conditions had very similar expected bonuses; however, the condition with only two levels in both the cause and the effect was unique in that it had expected bonuses at the high range of the other conditions. (This arises because even though standardizing the variables means that they technically have the same variance, the distribution of a binary variable is inherently different from the distribution of multi-level variables.) I was concerned that participants in this condition might obtain slightly higher than average bonuses, which could subsequently make them feel as if they performed better and influence their final judgments. This would be especially problematic for testing the feelings-of-success heuristic, as it also predicts stronger judgments when the effect has fewer levels.

Another concern with the equation above when the effect only had two levels was that it resulted in a bonus that was greater than zero if the prediction was incorrect. It felt odd to have a non-zero bonus when choosing the wrong option out of only two options.

To fix both of these problems, I sought a bonus value for situations in which the effect only has two levels such that a wrong prediction would receive a bonus of zero points, and I examined different possible bonuses for a correct prediction. There was no way to get the distribution of the expected bonuses to match exactly across all conditions, but I eventually settled on 8.5 points for a correct prediction. This bonus produced expected total bonuses for conditions in which $\text{Levels}_E = 2$ that were close to, but slightly lower than for conditions in which $\text{Levels}_E = 5$ and 13. This bonus now sets up a conservative test of the feelings-of-success heuristic.

In summary, the bonus scheme was set up to encourage participants to adopt a metric of accuracy analogous to how linear regression works, and was designed so that the average bonuses participants obtained should be roughly equivalent across all conditions. Because it was not possible to make the average bonuses exactly the same across conditions, the bonuses for conditions in which the effect had two levels were slightly lower to set up a conservative test of the feelings-of-success heuristic. Although it would be ideal if the bonuses could be viewed as a true and objective measure of accuracy in the prediction task, for all the reasons mentioned above it appears to be very challenging to have a perfect measure of prediction accuracy that is equivalent across conditions; bonuses can be viewed as a rough measure of accuracy.

2.1.6 Measures of causal strength

After making predictions on all trials for a given scenario, participants made judgments concerning the entire dataset they had just experienced. Each participant made judgments using one of the three measures discussed in Section 1.4. Table 3 displays the prompts and response scales corresponding to each of these three measures.

Table 3. ***Measures Used in Experiments***

Construct	Measure	Prompt	Level and label	
Objective strength	*Causal strength*	What is the causal relationship between [drug] and [enzyme]?	5	High levels of [drug] strongly cause low levels of [enzyme]
			0	No causal relationship
			-5*	High levels of [drug] strongly cause high levels of [enzyme]
	Predictiveness	How well does the dosage of [drug] predict the level of [enzyme]?	7	Very predictive
			1	Not predictive
Subjective feelings-of-success	*Prediction success*	How successful were you in predicting the level of [enzyme]?	7	Very successful
			1	Not successful

Note. Within each scenario, the name of a drug is displayed in place of [drug] and the name of an enzyme is displayed in place of [enzyme]. *All analyses in this paper use the absolute value of the causal strength measure so that they can be easily compared to analyses of the other two measures.

2.2 Results

The outline of the results section is as follows. First, I will present data pertaining to participants' accuracy on the trial-by-trial prediction across conditions. Second, I test if participants' judgments are based on the specificity criterion, the subjective feelings-of-success heuristic, or the proportionality criterion. To do so, I analyzed participants' judgments on each of the measures, to determine if there were influences from the granularities of the cause and effect, as well as differences between proportional vs. non-proportional conditions. I conducted these analyses both with and without controlling for the bonus points participants won on a particular scenario.

2.2.1 Prediction accuracy

How well participants performed in the trial-by-trial prediction task was not the main focus of the experiment, but was measured for two reasons. First, I wanted to ensure that participants were properly learning about the causal relationships in the datasets. If they were attending to the observations and learning more about the relationship as they experienced more trials, their predictions should get more accurate with increasing trials. I reported participants' trial-level prediction accuracy in Appendix B. Participants appeared to be increasingly accurate in predicting the state of the effect from the cause over time.

The second reason to measure prediction accuracy was to ensure the difficulty of the prediction task was comparable across conditions with different granularities in the cause and effect. The accuracy bonus scoring scheme was designed such that participants should obtain roughly equal accuracy (and bonuses) across all conditions, and if there is any difference, that participants would obtain higher bonuses for conditions with more levels in the effect. The total bonus points won by participants for each scenario are displayed in Figure 4.

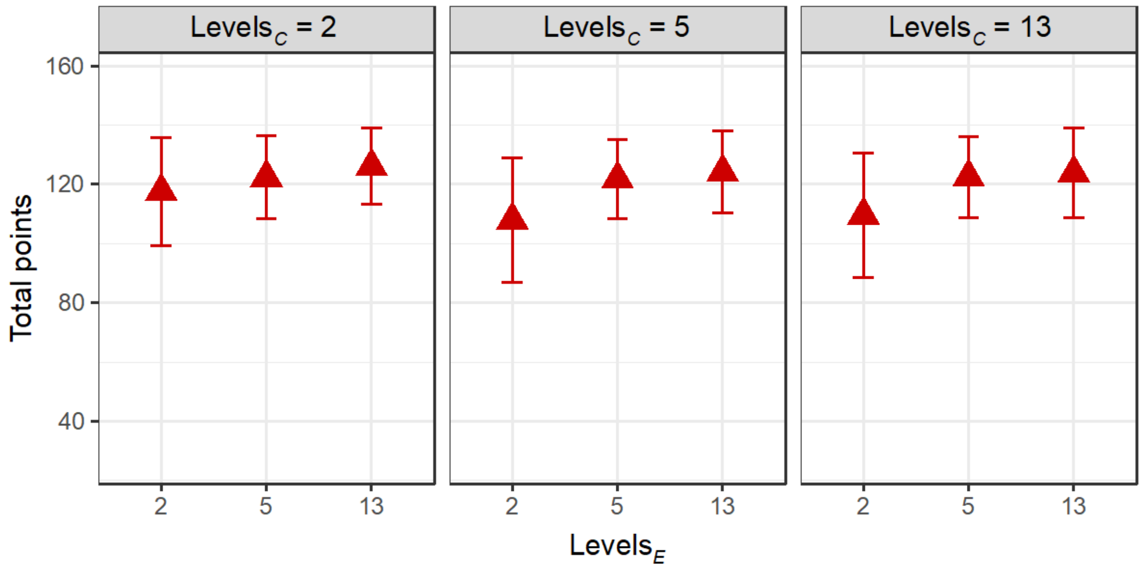

***Figure 4*. Total bonus points won by participants in each scenario by the granularity of the cause and effect in Experiment 1. Triangles represent condition means, and error bars represent standard deviations. Total bonus points won for conditions in which Levels$_E$ = 2 are clustered at various levels because participants are awarded either zero or 8.50 points for each prediction, leading to total points in multiples of 8.50.**

To assess if there were differences in prediction accuracy across conditions, I ran a regression predicting the total bonus points won in each scenario from Levels$_C$ and Levels$_E$, with by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures. There was a small but marginally significant effect of Levels$_C$ (B = -0.22, 95% CI [-0.44, -0.01], p = .042, R^2_{NSJ} = .004, 95% CI [.000, .015]); participants scored slightly more points when the cause had fewer levels. There were significant differences in total points across conditions with different Levels$_E$ (B = 1.00, 95% CI [0.77, 1.22], p < .001, R^2_{NSJ} = .072, 95% CI [.045, .103]); participants scored more points in scenarios in which the effect had more levels.

Ideally, participants would have obtained comparable bonus totals across all conditions. This was not achieved, but it was not problematic because the results showed that participants scored fewer points when the effect had fewer levels, which worked against the feelings-of-success hypothesis and provided a more conservative test when analyzing the influence of the granularity of the effect. However, since the analysis above revealed differences across conditions, when testing the influence of the granularities of the cause and effect, I performed analyses of the raw judgments, and also repeated the analyses after controlling for the bonus points won within each scenario.

2.2.2 Testing the specificity criterion, the feelings-of-success heuristic, and the proportionality criterion

If participants were sensitive to the specificity criterion, they would give stronger judgments in conditions with more levels in the cause (a positive effect of $Levels_C$). If participants used the feelings-of-success heuristic, they would give stronger judgments in conditions with fewer levels of the effect (a negative effect of $Levels_E$). If participants were sensitive to the proportionality criterion, they would give stronger judgments in conditions in which the cause and effect had the same granularity compared to conditions in which they had differing granularities (a positive effect of proportionality).

All regressions were run using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). I ran separate regressions for each of the measures (*causal strength*, *predictiveness*, and *prediction success*) predicting each from the granularities of the cause and effect ($Levels_C$ and $Levels_E$), and a grouping variable indicating whether a particular condition had proportional vs. non-proportional causes and effects. I treated $Levels_C$ and $Levels_E$ as numerical predictors, since I was interested in the influence of each along the whole (theoretical) scale of granularity. I treated proportionality as a factor using effect codes of -0.5 and +0.5 for each condition.

In each regression, I also controlled for the objective causal strength (either $r = .60$ or -.60) by including it as a predictor, and included by-subject random slopes for Levels_C, Levels_E, and proportionality. The maximal regression model predicting the measure of *prediction success* failed to converge, so I dropped the correlation between the random slope of proportionality with the random slopes for Levels_C and Levels_E (see Barr, Levy, Scheepers, & Tily, 2013).

I ran a second set of regressions that were identical to the first set, but controlled for the total bonus points won in each scenario. In this set, the maximal regression models for the *causal strength* and *prediction success* measures failed to converge, so I again dropped the correlation between the random slope of proportionality and the other predictors. The results for both sets of regressions are presented in Table 4.

Table 4. ***Multivariate Regression Results for Experiment 1***

Predictor	Not controlling for bonus points			Controlling for bonus points		
	B	p	R^2_{NSJ}	B	p	R^2_{NSJ}
Measure: *Causal strength*						
Levels_C	-0.01 [-0.04, 0.05]	.828	.000 [.000, .015]	0.02 [-0.03, 0.06]	.424	.002 [.000, .021]
Levels_E	-0.11 [-0.15, -0.07]	< .001	.058 [.020, .113]	-0.13 [-0.18, -0.09]	< .001	.091 [.043, .154]
Proportionality	0.07 [-0.34, 0.48]	.742	.000 [.000, .016]	0.00 [-0.40, 0.40]	.985	.000 [.000, .014]
Measure: *Predictiveness*						
Levels_C	-0.01 [-0.02, 0.04]	.516	.001 [.000, .018]	0.01 [-0.02, 0.04]	.440	.001 [.000, .019]
Levels_E	-0.10 [-0.14, -0.07]	< .001	.097 [.047, .159]	-0.13 [-0.16, -0.10]	< .001	.132 [.075, .200]
Proportionality	0.39 [0.10, 0.69]	.011	.015 [.001, .050]	0.34 [0.06, 0.62]	.018	.012 [.000, .044]
Measure: *Prediction success*						
Levels_C	0.01 [-0.02, 0.03]	.525	.001 [.000, .017]	0.02 [-0.01, 0.04]	.131	.004 [.000, .028]
Levels_E	-0.02 [-0.05, 0.01]	.206	.005 [.000, .030]	-0.07 [-0.10, -0.04]	< .001	.049 [.015, .099]
Proportionality	0.25 [0.01, 0.50]	.047	.007 [.000, .035]	0.14 [-0.09, 0.36]	.241	.002 [.000, .023]

Note. Confidence intervals represent 95% CIs. The reported effects size measure (R^2_{NSJ}) is analogous to partial-R^2 in that it captures the conditional variance explained by each predictor, specifically for generalized linear mixed models (Jaeger, Edwards, Das, & Sen, 2017).

Participants' judgments are plotted in Figures 5 and 6, using the visreg package in R (Breheny & Burchett, 2017). Given the random effects structure of the models, the 95% confidence intervals had to be computed via simulation using the bootpredictlme4 package in R (Duursma, 2017). The data in Figures 5 and 6 actually represent judgments after controlling for a set of predictors in the model. In Figure 5, the judgments are plotted after controlling for predictors other than Levels_C and Levels_E, so that the effect of the two predictors can be visualized. Figure 6 is similar, except that judgments are plotted after controlling for predictors other than the grouping variable indicating if a condition was proportional or non-proportional.

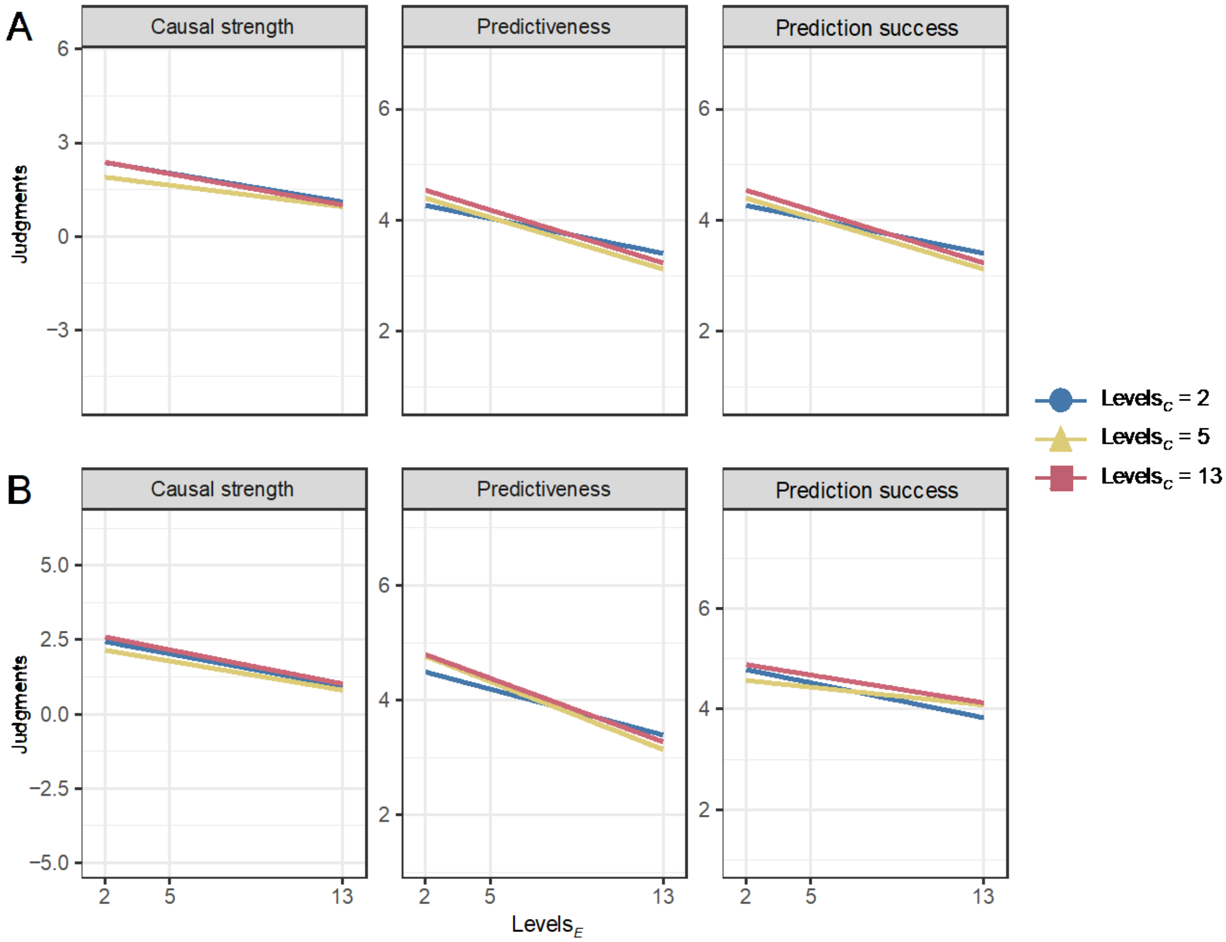

***Figure 5*. Participants' judgments by Levels$_C$ and Levels$_E$ after controlling for other predictors in the two sets of regression models in Experiment 1. Ribbons indicate 95% CIs. (A) Judgments without controlling for bonus points. (B) Judgments after controlling for bonus points.**

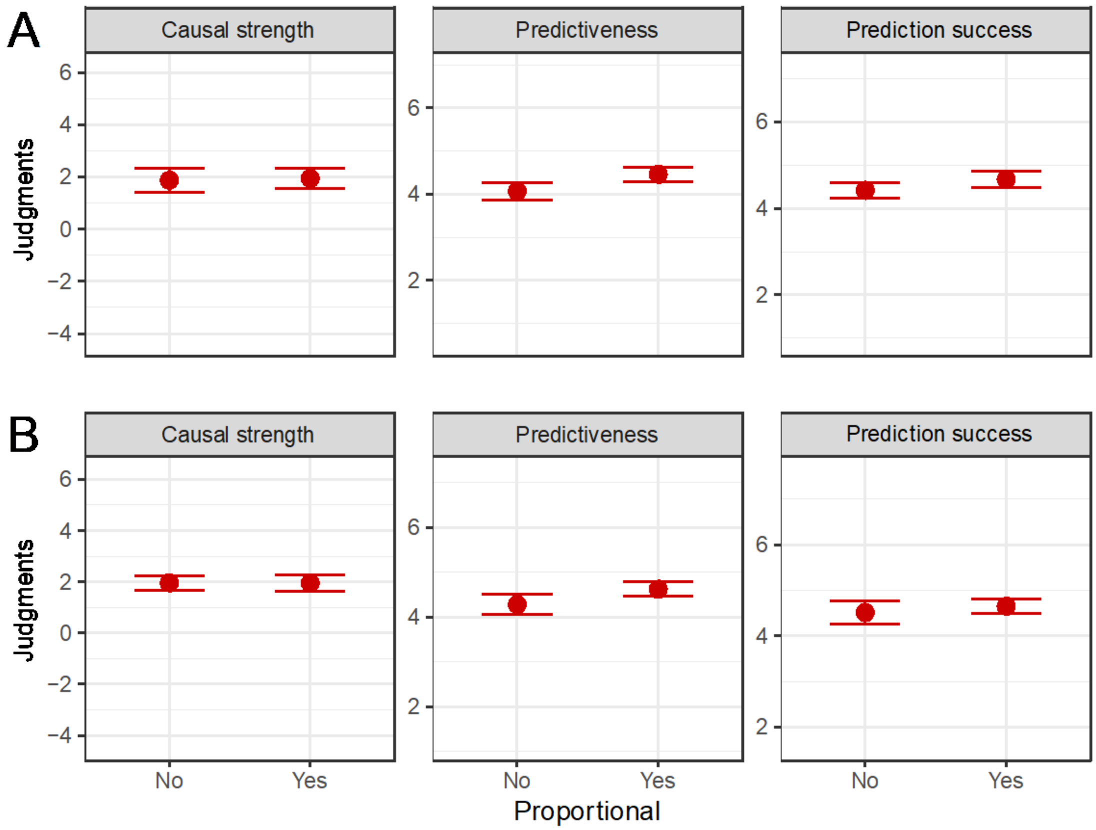

***Figure 6*. Participants' judgments for proportional vs. non-proportional conditions after controlling for other predictors in the two sets of regression models in Experiment 1. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments without controlling for bonus points. (B) Judgments after controlling for bonus points.**

The granularity of the cause (Levels$_C$) was not a significant predictor of any of the three measures. In Figure 5, there are no differences in the lines representing each condition of Levels$_C$. Thus, participants were not sensitive to the specificity criterion.

In contrast, the granularity of the effect (Levels$_E$) was a significant predictor of *causal strength* and *predictiveness*, both before and after controlling for the bonus points won on each scenario. The granularity of the effect was also a significant predictor for judgments of *prediction success*, but only after controlling for the bonus points won. Participants made weaker judgments when the effect had more levels, as implied by the negative coefficients for Levels_E in Table 4 and the decreasing slopes in Figure 5.

I had initially predicted that the granularity of the effect would likely have an influence on judgments of *prediction success*, which measured the subjective feelings participants experienced when making predictions. However, the fact that the granularity of the effect also had an influence on judgments of *causal strength* and *predictiveness* was more surprising, as those dependent variables measured participants' assessments of the objective causal strength in the data. This finding provides some initial evidence that people use their subjective feelings-of-success for estimating the strength of the relation between a cause and effect.

Judgments of *predictiveness* were slightly greater in proportional vs. non-proportional conditions. However, there were no differences between these conditions in judgments of *causal strength* or *prediction success.* In sum, participants appear to be partly sensitive to the proportionality criterion when judging the *predictiveness* of a cause, but not when making other judgments.

2.3 Discussion

More levels in the effect (Levels_E) led to weaker judgments of *causal strength*, *predictiveness*, and *prediction success*. It was relatively unsurprising that the *prediction success* measure, which measured participants' subjective feelings-of-success, was influenced by the granularity of the effect; participants may have found the prediction task more difficult (and perceived themselves to be less successful) when making predictions of more fine-grained effects because they were less likely to predict the exact true state of the effect. However, the measures of *causal strength* and *predictiveness* both assessed the objective strength of the causal relationship in each scenario. Despite the objective strength across all conditions being fixed at $r = \pm .60$, participants perceived the objective strength to be stronger when the effect had fewer levels.

The influence of the effect's granularity on all measures is consistent with participants making judgments using the feelings-of-success heuristic. Instead of directly assessing the objective strength of the causal relationship, participants appear to be making judgments based on their experience of using the cause to predict the effect. In other words, participants' subjective feelings-of-success may be bleeding over to influence their perceptions of the objective causal relationship.

The idea that subjective feelings-of-success may influence judgments of objective causal strength bears some resemblance to findings that people use subjective feelings as cues to various kinds of judgments about more objective quantities (Greifeneder, Bless, & Pham, 2011; Oppenheimer, 2008). For example, when people learn new information, their feelings of fluency (how easy they felt it was to learn the information) can influence their judgments of how well they have actually learned that information (Hertzog et al., 2003; Koriat, 2008). As another example, people will defer choices about consumer products when descriptions about them are more difficult to read or understand, because the subjective feeling of difficulty when processing or learning the information is taken as a cue about the product's objective value (Novemsky, Dhar, Schwarz, & Simonson, 2007). Similarly, in the present experiment, people arrived at indirect estimates of a target quantity by substituting a secondary and more easily accessible quantity. The feeling of how successful one has been in the prediction task may be used as a cue or substitute for the objective strength of the causal relationship.

In contrast, the granularity of the cause (Levels$_C$) did not influence participants' judgments; participants appear insensitive to the specificity criterion, which states that causes that are more fine-grained – i.e. have more levels – will be judged as better (in the present context, stronger or more predictive).

A comparison of proportional vs. non-proportional conditions showed that participants were at least somewhat sensitive to the *proportionality* criterion; participants made stronger judgments using the *predictiveness* measure in conditions for which the cause and effect had the same number of levels. The proportionality criterion is a normative criterion for determining how good a causal explanation is based on its level of granularity (McGrath, 1998; Woodward, 2010; Yablo, 1992). In the present experimental task, the *predictiveness* measure can be viewed as an evaluation of how good or useful a cause is for the purpose of predicting the effect. This explains why there was an effect of proportionality in the *predictiveness* but not the *causal strength* measure (after controlling for bonus points). Despite both measures targeting the objective causal strength in the data, the *causal strength* measure elicits a judgment about the degree and polarity (positive vs. negative) of the relationship without focusing on the cause's utility for prediction.

In sum, the present experiment showed that when learning about a causal relationship, people are primarily sensitive to the granularity of the effect, and partly to the proportionality between the cause and effect. The findings are consistent with participants using the feelings-of-success heuristic, and of them being partly sensitive to the proportionality criterion.

3.0 Experiment 2: Do bonus points drive feelings-of-success?

The primary finding from Experiment 1 was that when learning about a causal relationship via a trial-by-trial prediction task, participants' judgments of *causal strength*, *predictiveness*, and *prediction success* were influenced by the granularity of the effect. Participants made weaker judgments when the effect had fewer levels, even though the objective causal strength was held constant. This finding was consistent with the hypothesis that participants made judgments based on their subjective feelings-of-success.

When making predictions of an effect with many levels, participants may have believed the prediction task was more difficult because they were less likely to predict the exact state of the effect. However, when predicting the state of a fine-grained effect, a prediction with small (as opposed to zero) error should count as an accurate prediction. I attempted to encourage participants to think about their prediction accuracy in this way by awarding them accuracy bonuses according to a scheme that would lead to similar bonus totals across conditions. The bonus points served as an external cue to prediction accuracy other than participants' subjective feelings-of-success. Even though the bonus points across conditions indicated that participants were more accurate when the effect was fine-grained, and even though I controlled for bonus points when analyzing participants' judgments, participants still made weaker judgments when the effect was more fine-grained.

A weakness of the design in Experiment 1 was that the bonus points that participants obtained in the prediction task varied to some extent across conditions due to both statistical issues with equating accuracy across conditions as well as potential differences in participants' learning (see the method and results sections about bonuses in Experiment 1 for details). Although in Experiment 1 the influence of the granularity of the effect held with and without controlling for bonus points, I wanted to verify that this finding was not produced by showing participants bonus points for each prediction. Learning and making predictions without being awarded bonus points is a better analog of learning in real-world settings.

Thus, in Experiment 2, I tested a condition in which participants were not shown bonus points for their predictions, and compared their judgments to a control condition in which they received bonus points for each prediction, similar to Experiment 1.

If the influence of the effect's granularity in Experiment 1 was driven by the fact that participants had an external cue about their prediction accuracy, then the findings from Experiment 1 might not replicate when the bonus points are not shown to participants. However, if the influence of the effect's granularity was indeed driven by the change in the subjective feelings-of-success during the prediction task, then the same findings should hold regardless of whether or not bonus points are shown to participants.

3.1 Method

3.1.1 Participants

I recruited 240 new participants on Mturk using the same criteria as in Experiment 1. Data from one participant was not recorded due to a programming error. Each participant was paid \$0.90, and an additional accuracy bonus (*M* = \$0.47, *SD* = \$0.04) based on their performance in the trial-by-trial prediction task. The experiment lasted between 6-8 minutes.

3.1.2 Design and procedure

The design of Experiment 2 was largely similar to the design of Experiment 1. There was a within-subject manipulation of the number of levels in the cause and effect. However, in this experiment, the granularities of the cause and the effect were manipulated so that each variable had either two or 13 levels (resulting in four within-subject conditions based on granularity). I eliminated the conditions in which the cause or effect had five levels because the results of Experiment 1 revealed the influence of the effect's granularity was strongest when comparing the conditions in which the effect had two levels vs. 13 levels. Thus, I retained the conditions that would result in the strongest comparison.

Experiment 2 had a new between-subjects factor; participants either received bonus points or no bonus points regarding the accuracy of their predictions on each trial. The bonus condition looked identical to Experiment 1, except that instead of displaying a running total of the bonus points accrued across the entire experiment, a running total of the bonus points accrued across the current scenario was displayed above the stimuli (Figure 7A, compare to Figure 3), while the overall total across all prior scenarios was displayed on the upper right corner of the experiment screen (it is not visible in Figure 7A). I made this change to further highlight participants' levels of prediction accuracy within each scenario.

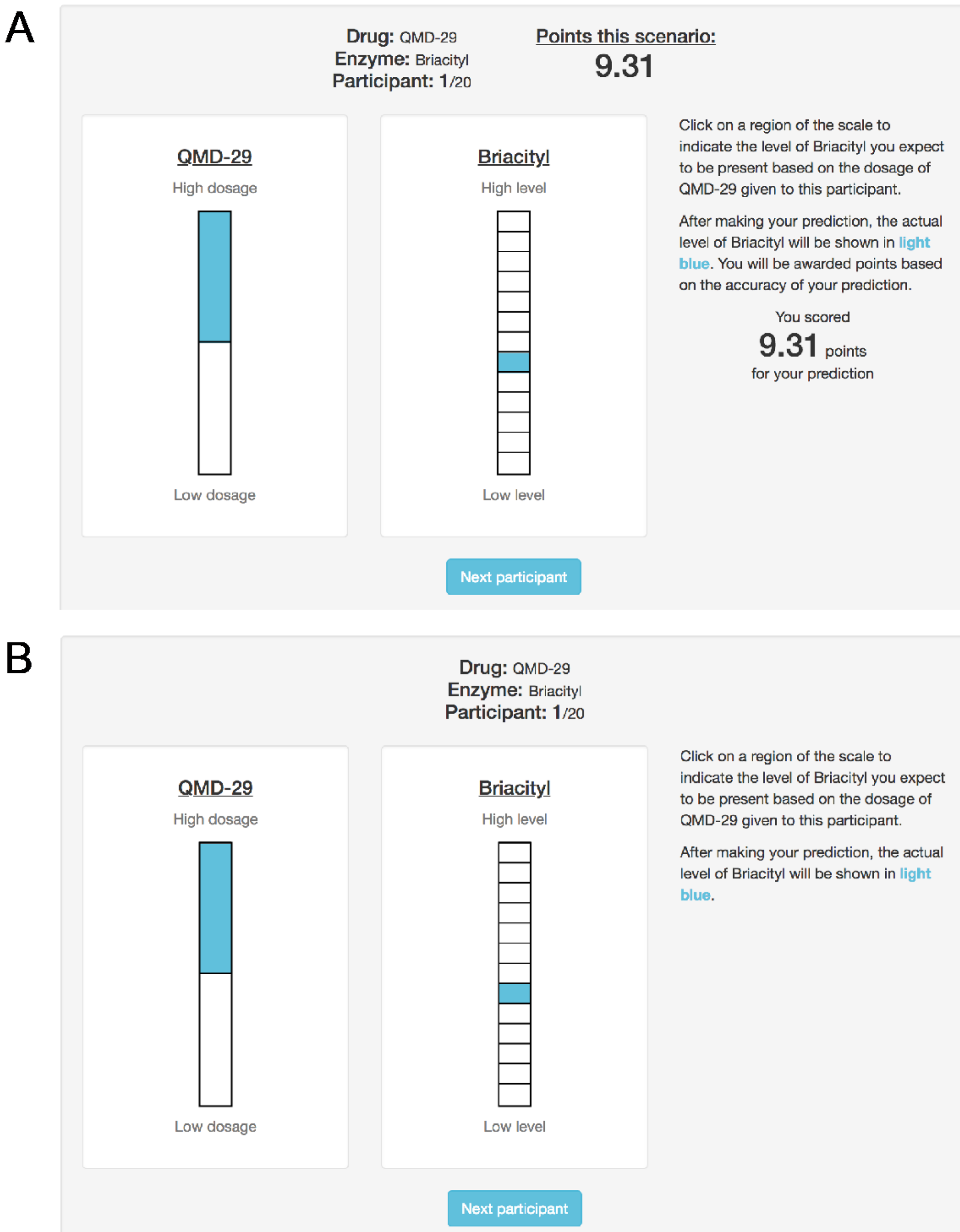

***Figure 7*. Presentation of stimuli in Experiment 2 in the bonus condition (A) and the no-bonus condition (B). These screen shots display what the screen looks like after a participant has made their prediction; the screen now shows the actual level of the effect (Briacityl).**

In the no-bonus condition, participants were not told about or shown any bonus points during the experiment (Figure 7B). However, the points participants would have won (based on the scoring scheme) for their predictions in the trial-by-trial prediction task were recorded, and participants were awarded a bonus payment based on their total bonus points at the end of the experiment for fairness.

Similar to Experiment 1, participants responded to one of the three measures at the end of each scenario, between-subjects. Thus, there were six between-subject conditions, with 40 participants assigned to each condition.

3.2 Results

The plan for analyzing data from Experiment 2 mirrored the analyses of Experiment 1, except that I also included results for the no-bonus condition. In Experiment 1, I treated Levels_C and Levels_E as numerical predictors; coefficients from regressions representing the strength of these effects represented changes in judgments due to single-unit changes in granularity. This made sense when considering how granularity across the entire scale could influence participants' judgments. However, in Experiment 2, both Levels_C and Levels_E had two levels (two vs. 13). In the present analyses, for ease of interpretation, I treated the effects of Levels_C and Levels_E as factors using effect codes of -0.5 and +0.5 for each level, so that coefficients from regressions represented differences between conditions.

3.2.1 Prediction accuracy

As in Experiment 1, participants' prediction accuracy was not the main focus of the analyses, but were measured to ensure participants were learning about the causal relationship from the data and to measure participants' performance in the prediction task across conditions.

The trial-level prediction accuracy of participants is reported in Appendix B. From the trial-level data, participants appeared to be increasingly accurate in predicting the state of the effect from the cause over time, suggesting that they were learning more about the causal relationship with more trials. Crucially, this was apparent both in the bonus and no-bonus condition, indicating that even without accuracy bonuses for their predictions, participants were still attentive to the task.

The bonus points won by participants for each scenario are displayed in Figure 8. (As a reminder, in the no-bonus conditions, participants were only alerted to their total bonus at the end of the study, not after each trial.)

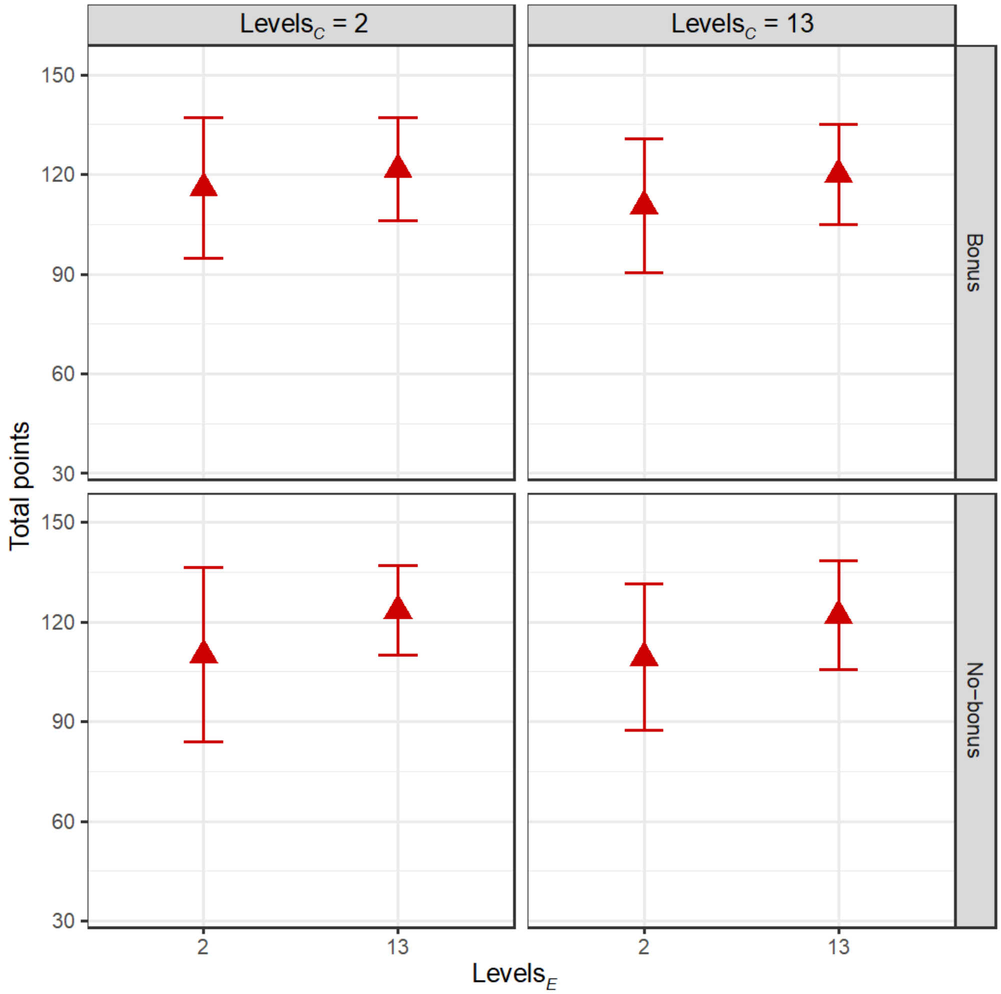

***Figure 8*. Bonus points won by participants in each scenario by the granularity of the cause and effect, and bonus condition in Experiment 2. Triangles represent means, and error bars represent standard deviations.**

To assess the differences in bonus point totals across different conditions, I ran a regression predicting the bonus points won in each scenario from Levels$_C$, Levels$_E$, the bonus condition, as well as two-way interactions between each of Levels$_C$ and Levels$_E$ with the bonus condition. The regression included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures. There was a small but marginally significant effect of Levels$_C$ (B = -3.47, 95% CI [-6.61, -0.36], p = .030, R^2_{NSJ} = .004, 95% CI [.000, .016]); participants scored more points when the cause had fewer levels. There were significant differences in total points across conditions with different Levels$_E$ (B = 7.56, 95% CI [3.96, 11.16], p < .001, R^2_{NSJ} = .019, 95% CI [.006, .040]); participants scored more points in scenarios in which the effect had more levels. The patterns of findings were similar to Experiment 1; by design, the scoring scheme helped ensure that participants scored fewer points in conditions in which the effect had fewer levels (Levels$_E$ = 2), which provided a conservative test of the feelings-of-success hypothesis when analyzing the influence of the granularity of the effect.

There were no differences in bonus points between the bonus vs. no-bonus conditions (p = .589), and no interaction between Levels$_C$ and the bonus factor (p = .302). However, there was an interaction between Levels$_E$ and bonus condition; the influence of Levels$_E$ was greater amongst participants in the no-bonus condition (B = 5.45, 95% CI [0.35, 10.55], p = .037, R^2_{NSJ} = .005, 95% CI [.000, .018]). Although I did not have specific a priori hypotheses about this, it appears from Figure 8 that participants in the no-bonus condition had lower bonuses when the effect only had two levels. It is possible that the zero point bonuses when they were wrong in the bonus condition spurred them to learn and make predictions a bit better.

Since the analysis above revealed differences across conditions, when testing the influence of the granularities of the cause and effect, I performed analyses of the raw judgments, and for the bonus condition also repeated the analyses after controlling for the bonus points won within each scenario.

3.2.2 Testing the specificity criterion, the feelings-of-success heuristic, and the proportionality criterion

In Experiment 1, participants' judgments were sensitive to the granularity of the effect (Levels$_E$), consistent with them using the feelings-of-success heuristic. However, they were not sensitive to the granularity of the cause (Levels$_C$), implying they were not sensitive to the specificity criterion. There were differences in judgments of *predictiveness* between proportional and non-proportional conditions but not the other measures, indicating partial sensitivity to the proportionality criterion. The main purpose of the analyses in this section was to determine if this pattern of findings would be replicated even when participants were not shown any bonus points during the prediction task.

For the bonus and no-bonus conditions separately, I ran regressions for each of the measures (*causal strength*, *predictiveness*, and *prediction success*), predicting each from Levels$_C$ and Levels$_E$ (treated as factors), and a grouping variable indicating whether a particular condition had proportional vs. non-proportional causes and effects. The regressions controlled for the objective causal strength of the datasets, and included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures.[7]

For the regressions in the bonus condition, I ran accompanying regressions that were identical to the first set, but also controlled for the bonus points won in each scenario. It did not make sense to run regressions controlling for the bonus points won per scenario in the no-bonus condition, because participants were never shown bonus points as an external cue to their prediction accuracy.

The results of these regressions are presented in Table 5. Figure 9 displays the effects of Levels$_C$ and Levels$_E$ on participants' judgments after controlling for all the other predictors in each

---

[7] For the regression analyses in Experiment 2 and subsequent experiments, I did not include a by-subject random slope for proportionality in the regression models because there were insufficient observations to estimate a random slope for proportionality in addition to Levels$_C$ and Levels$_E$. In Experiment 2 and subsequent experiments, there were four scenarios per participant. In Experiment 1, there were nine scenarios per participant, which was sufficient to estimate the three by-subject random slopes.

regression. Figure 10 displays participants' judgments after controlling for all predictors other than the grouping variable indicating if a condition was proportional or non-proportional.

Table 5. ***Multivariate Regression Results for Experiment 2***

Predictor	No-bonus condition			Bonus condition			Bonus condition (controlling for bonus points)		
	B	p	R^2_{NSJ}	B	p	R^2_{NSJ}	B	p	R^2_{NSJ}
Measure: *Causal strength*									
$Levels_C$	-0.19 [-0.69, 0.31]	.472	.002 [.000, .040]	0.08 [-0.55, 0.70]	.812	.000 [.000, .033]	0.19 [-0.40, 0.78]	.534	.002 [.000, .040]
$Levels_E$	-0.92 [-1.48, -0.36]	.002	.049 [.005, .131]	-0.87 [-1.56, -0.18]	.017	.035 [.001, .110]	-1.27 [-1.93, -0.62]	< .001	.080 [.019, .174]
Prop.	-0.56 [-1.03, -0.08]	.023	.019 [.000, .081]	-0.06 [-0.66, 0.54]	.851	.000 [.000, .032]	-0.23 [-0.78, 0.32]	.421	.003 [.000, .043]
Measure: *Predictiveness*									
$Levels_C$	0.08 [-0.29, 0.45]	.675	.001 [.000, .035]	-0.20 [-0.57, 0.17]	.294	.004 [.000, .048]	0.00 [-0.32, 0.32]	.987	.000 [.000, .031]
$Levels_E$	-1.32 [-1.79, -0.85]	< .001	.176 [.085, .287]	-0.94 [-1.43, -0.45]	< .001	.081 [.020, .176]	-1.34 [-1.74, -0.93]	< .001	.192 [.098, .304]
Prop.	0.30 [-0.05, 0.65]	.094	.011 [.000, .066]	-0.29 [-0.65, 0.08]	.126	.009 [.000, .060]	-0.17 [-0.42, 0.09]	.201	.004 [.000, .047]
Measure: *Prediction success*									
$Levels_C$	-0.26 [-0.59, 0.07]	.152	.007 [.000, .056]	-0.29 [-0.69, 0.11]	.157	.012 [.000, .068]	-0.20 [-0.51, 0.11]	.216	.007 [.000, .056]
$Levels_E$	-1.73 [-2.14, -1.32]	< .001	.268 [.163, .382]	-0.77 [-1.18, -0.36]	< .001	.083 [.021, .178]	-1.07 [-1.46, -0.68]	< .001	.173 [.082, .283]
Prop.	0.37 [0.08, 0.66]	.017	.016 [.000, .077]	-0.03 [-0.41, 0.35]	.876	.000 [.000, .032]	-0.21 [-0.50, 0.08]	.152	.009 [.000, .059]

Note. Confidence intervals represent 95% CIs. Prop. = Proportionality.

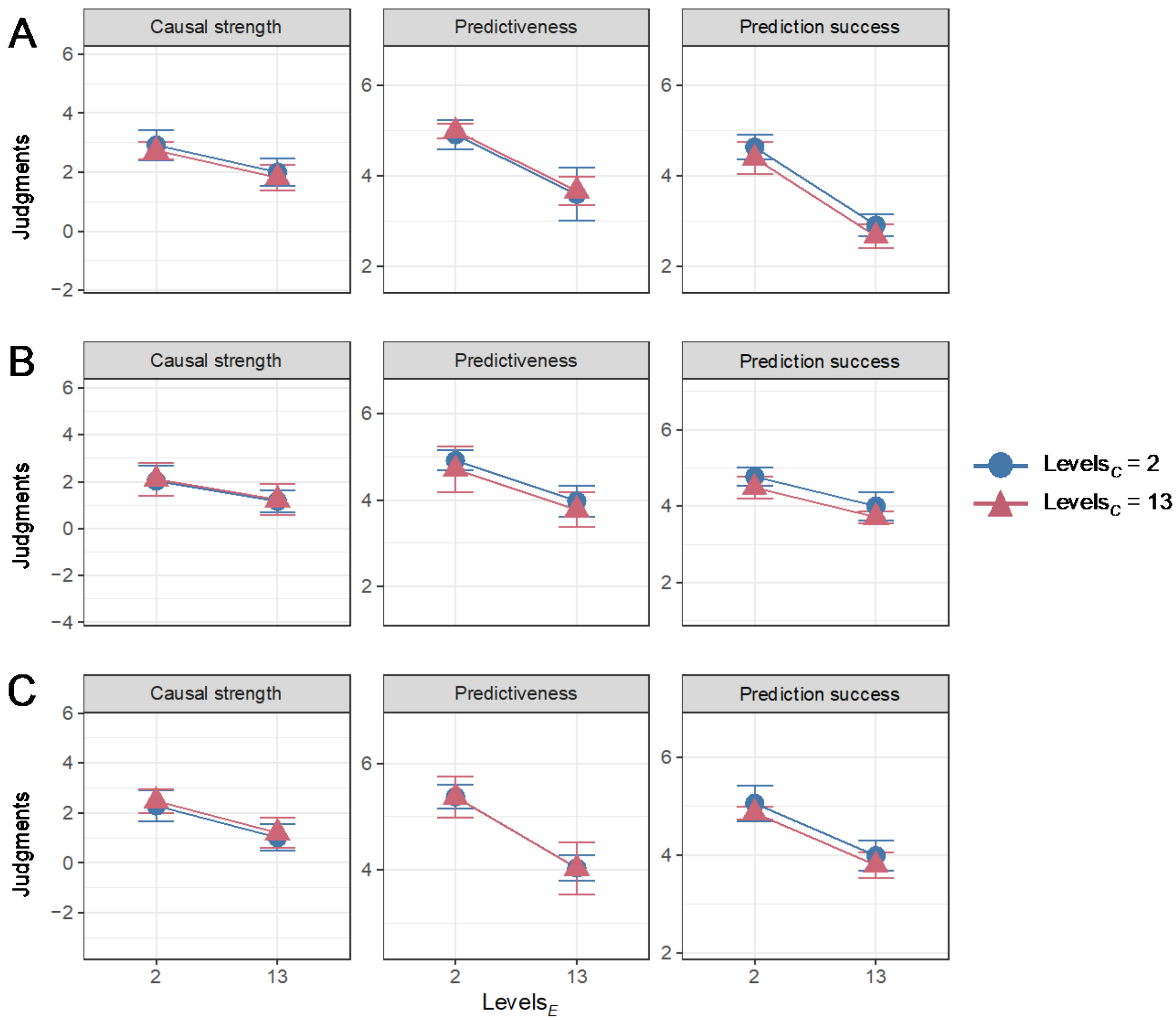

***Figure 9*. Participants' judgments by Levels$_C$ and Levels$_E$ after controlling for other predictors in the regression models in Experiment 2. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments from no-bonus condition without controlling for bonus points. (B) Judgments from bonus condition without controlling for bonus points. (C) Judgments from bonus condition after controlling for bonus points.**

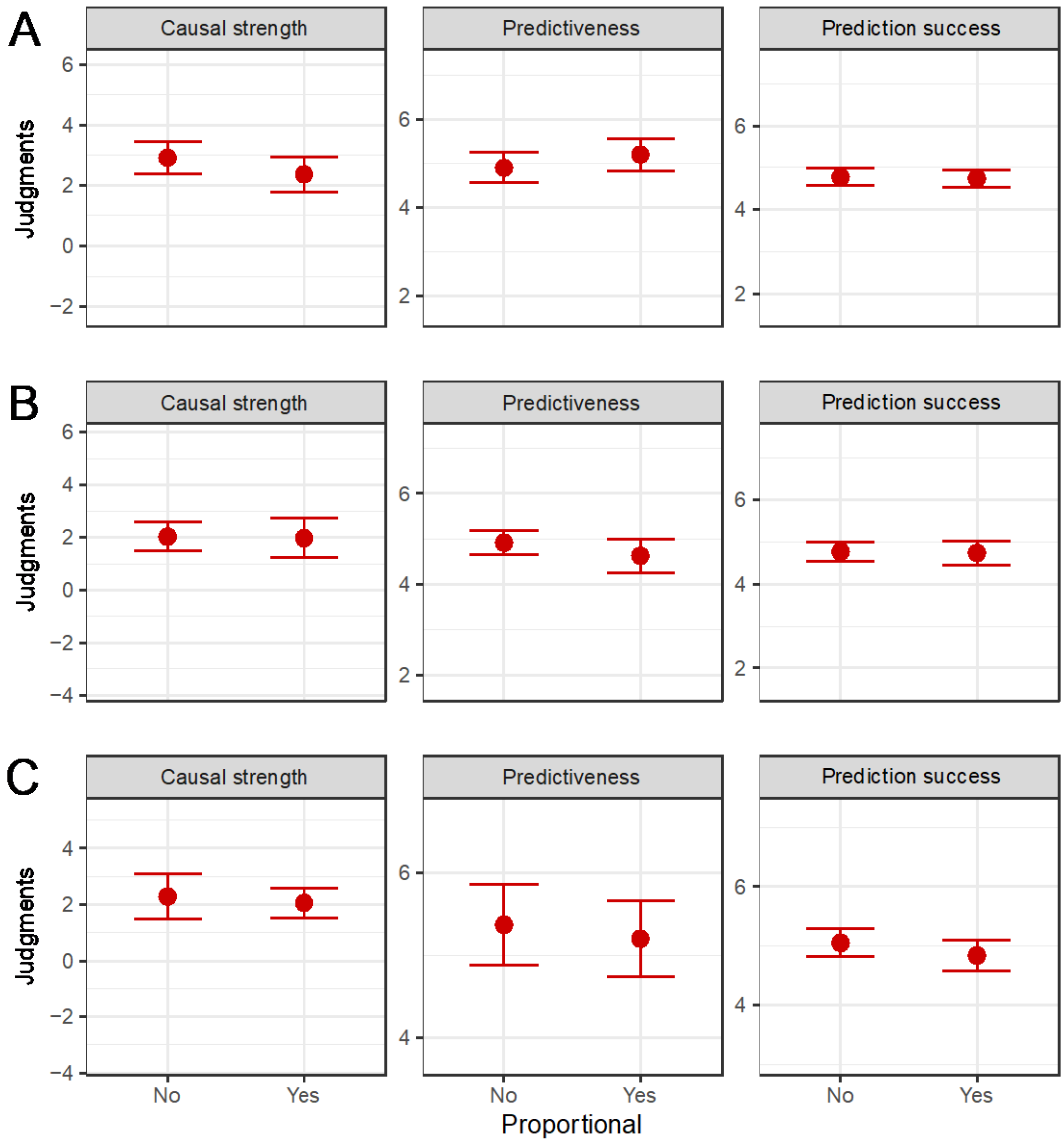

Figure 10. **Participants' judgments for proportional vs. non-proportional conditions after controlling for other predictors in the regression models in Experiment 2. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments from no-bonus condition without controlling for bonus points. (B) Judgments from bonus condition without controlling for bonus points. (C) Judgments from bonus condition after controlling for bonus points.**

3.2.2.1 The influence of Levels$_C$

Similar to Experiment 1, the granularity of the cause was not a significant predictor of any judgments. Crucially, this finding was consistent across both the no-bonus and bonus conditions (even after controlling for the total bonus points). Thus, participants were not sensitive to the specificity criterion.

3.2.2.2 The influence of Levels$_E$

The granularity of the effect was a significant predictor of judgments on all measures. This finding was consistent across both the no-bonus and bonus conditions, and also after controlling for the total bonus points per scenario in the bonus condition. Participants made weaker judgments when the effect had more levels, as implied by the negative coefficients for Levels$_E$ in Table 5 and the decreasing slopes in Figure 9. This result was consistent with participants using the feelings-of-success heuristic to make judgments about the data. Furthermore, since this result held even when participants did not receive any bonus points for their prediction accuracy, the result appears to be driven by participants' subjective feelings of prediction accuracy resulting from the granularity of the effect, rather than any external cue about their prediction accuracy.

Although the main purpose of Experiment 2 was to see if the influence of Levels$_E$ would be replicated even when participants did not receive any bonus points, I also tested whether receiving a bonus after each prediction moderated the influence of Levels$_E$. I ran separate regressions for each measure using data from both conditions, predicting judgments of each of the measures from Levels$_C$, Levels$_E$, the grouping variable indicating whether a particular condition had proportional vs. non-proportional causes and effects, the bonus condition, and the two-way interaction between Levels$_E$ and the bonus condition. The regressions included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures, and controlled for the objective causal strength of the datasets. The interaction term in these regressions compared the influence of Levels$_E$ in the bonus vs. no-bonus conditions (analogous to comparing rows A and B in Figure 9).

There was no interaction between the granularity of the effect and the bonus condition for judgments of *causal strength* (p = .949) or *predictiveness* (p = .399). However, the influence of the effect's granularity on judgments of *prediction success* was more strongly negative when participants were not shown bonus points (B = -0.95, 95% CI [-1.53, -0.38], p = .002, R^2_{NSJ} = .030, 95% CI [.004, .077]). This finding suggested that the accuracy bonus scoring scheme was working as intended. The scoring scheme was designed to make the prediction task appear equally difficult regardless of the cause and effect's granularity; without being shown bonus points, participants' subjective feelings-of-success were even more susceptible to differences in the effect's granularity.

3.2.2.3 The influence of proportionality

In the bonus condition, there were no differences in judgments between proportional vs. non-proportional conditions for any of the measures, both before and after controlling for bonus points. There were mixed findings regarding the effect of proportionality in the no-bonus condition. In line with the proportionality hypothesis, judgments of *prediction success* were significantly higher and judgments of *predictiveness* were marginally higher in the proportional condition. However, contradicting this hypothesis, judgments of *causal strength* were lower in the proportional condition.

In summary, the findings regarding proportionality are not reliable across Experiment 1 and Experiment 2, and they are also not reliable across the bonus vs. no-bonus conditions in Experiment 2. Further, it is not clear what is driving these differences, especially across experiments. Given the inconsistency of findings, I did not perform analyses to investigate the interaction between differences in the proportional vs. non-proportional conditions and the bonus condition.

3.3 Discussion

Consistent with the findings in Experiment 1, finer granularity in the effect led to weaker judgments of *causal strength*, *predictiveness*, and *prediction success*. Crucially, this finding was replicated both with and without bonuses.

When bonus points were not shown, the influence of the effect's granularity on judgments of *prediction success* became more extreme. This was expected; the purpose of the accuracy bonus was to equalize the difficulty of the prediction task across conditions. Without bonus points, participants did not have any external cue regarding their accuracy in the prediction task, leading to their subjective feelings-of-success being even more susceptible to differences in the effect's granularity. In other words, the consistent findings across the bonus and no-bonus conditions implied that participants' use of the feelings-of-success heuristic was not driven by any potential differences in accuracy bonuses.

So far, across Experiments 1 and 2, participants reliably gave stronger judgments when the effect was more coarse-grained, suggesting that participants were using the feelings-of-success heuristic. It is also clear that participants are not sensitive to the granularity of the cause, implying that they do not use the specificity criterion. The experiments have been inconsistent with regards to the proportionality criterion. Because the feelings-of-success heuristic seems to be by far the most reliable effect, the rest of the experiments largely focus on it.

When considering the feelings-of-success heuristic, the most obvious prediction is that participants would give stronger judgments of *prediction success* when the effect is more coarse-grained because they are more likely to get the prediction *exactly* correct. It is less obvious why the granularity of the effect has an influence on the *causal strength* and *predictiveness* measures; normatively, *causal strength* and *predictiveness* should only be influenced by the strength of the correlation between the cause and effect, which is held constant.

I am proposing that a fuller understanding of the feelings-of-success heuristic involves participants using their subjective feelings (measured by the *prediction success* measure) to make judgments of *causal strength* and *predictiveness*. Note that in Experiments 1 and 2, participants only made one of the three judgments, so the theory is that participants form feelings about *prediction success* even when they are not asked this question. In Experiment 3, I tested this hypothesis with a mediation analysis.

4.0 Experiment 3: A mechanism-level account for how granularity influences objective causal strength judgments

The primary finding in Experiments 1 and 2 was that participants' judgments of objective strength and subjective feelings-of-success were weaker when the effect had more levels. When the effect had more levels (i.e. was more fine-grained), it was subjectively more difficult to predict the effect given the cause, which is consistent with participant' ratings of subjective feelings-of-success (using the *prediction success* measure). However, normatively, the varying perceived levels of difficulty in the trial-by-trial task has no influence on the objective strength of the relationship. The goal of the present experiment was to uncover how the granularity of the effect comes to influence participants' judgments of objective strength (captured by the *causal strength* and *predictiveness* measures).

One hypothesis, as suggested in the discussion of Experiment 2, is that participants may be making indirect estimates of the objective causal strength by substituting a more accessible quantity like subjective feelings-of-success, which is influenced by the number of levels in the effect. In Experiment 2, the same influence of the number of levels in the effect on participants' judgments was found to hold whether or not participants received an accuracy bonus, indicating that participants relied on subjective feelings-of-success regardless of whether there was an external cue of their success in the prediction task.

Based on this hypothesis, Figure 11 presents a conceptual model for a potential mechanism by which the number of levels in the effect influences judgments about objective causal strength. In this model, subjective feelings-of-success is a potential mediator between the granularity of the effect and judgments of objective causal strength.

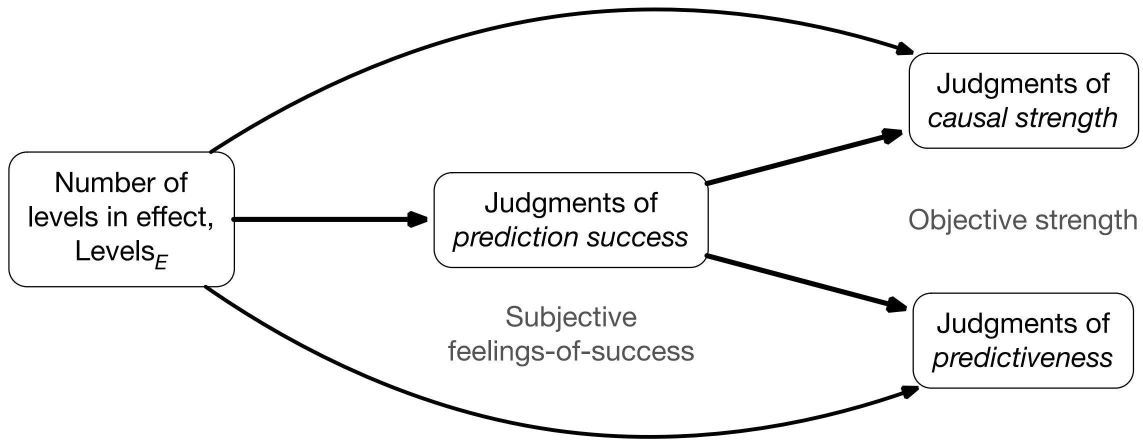

Figure 11. **Conceptual model for mediation relationship tested in Experiment 3.**

The granularity of the effect influences subjective feelings-of-success because it may appear easier to make more accurate predictions when the effect has fewer levels; participants are likelier to guess the correct state of the effect by chance when it has two levels compared to 13 levels. The measure of *prediction success* is intended to capture the subjective feelings-of-success that participants experience when they engage in the prediction task. It is likely that feelings-of-success arise spontaneously as participants make predictions, even without making judgments of *prediction success*. If they instead are forced to make judgments of the objective strength, they may substitute their subjective feelings-of-success from the prediction task instead of properly assessing the objective strength.

In Experiments 1 and 2, subjective feelings-of-success may be viewed as a reasonable substitute for objective causal strength if one assumes that stronger causes can be used to more accurately predict the effect. Research in other domains have demonstrated that people perform analogous “substitutions” of target quantities using more affective and subjective quantities (see

Gigerenzer & Goldstein, 2011; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Kahneman & Tversky, 1973; Koriat, 2008; Oppenheimer, 2008). The hypothesis in the present context is that participants substitute subjective feelings-of-success for their assessments of objective strength. In other words, the granularity of the effect may influence judgments of objective strength via subjective feelings-of-success as a mechanism.

The prior experiments were not designed to examine if subjective feelings-of-success was a mechanism linking the granularity of the effect with judgments about objective causal strength. In Experiments 1 and 2, participants made judgments using only one of the three measures, so a mediation analysis could not be performed. In the present experiment, I tested the mechanistic model in Figure 11 by having participants make judgments on each scenario using all three measures. If the model in Figure 11 is correct, judgments of *prediction success* should mediate the relationship between the granularity of the effect and judgments of *causal strength* and *predictiveness*.

4.1 Method

4.1.1 Participants

160 new participants were recruited on Mturk using the same criteria as in the prior experiments. Each participant was paid \$0.90, and an additional accuracy bonus (*M* = \$0.47, *SD* = \$0.04) for their performance in the trial-by-trial prediction task. The experiment lasted between six to eight minutes.

4.1.2 Design and procedure

The design of Experiment 3 was largely similar to Experiment 2. There was a within-subject manipulation of the number of levels in the cause (Levels$_C$) and effect (Levels$_E$); each variable had either two or 13 levels (resulting in four within-subject conditions).

In contrast to the prior experiments, in which participants made judgments using only a single measure, participants in Experiment 3 made judgments for all three measures (Table 3) at the end of each scenario. After the last trial, the first measure became visible below the stimuli. Each subsequent measure appeared in turn after participants submitted a judgment for the prior measure (they were not permitted to change their prior judgments after viewing the next measure). The order of the measures was partially counterbalanced between-subjects. Of the 160 participants, 40 experienced the measures in each of the four orders shown in Table 6.

Table 6. ***Order of Measures in Experiment 3***

	Order A	Order B	Order C	Order D
1	*Causal strength*	*Predictiveness*	*Predictive success*	*Predictive success*
2	*Predictiveness*	*Causal strength*	*Causal strength*	*Predictiveness*
3	*Predictive success*	*Predictive success*	*Predictiveness*	*Causal strength*

Different orders were used to ensure that any potential mediating effect was robust and not due only to an order effect. For example, if participants were always asked about the objective causal strength first, it is possible this would influence their subsequent judgments on the *prediction success* measure (rather than participants responding to it based on their experience in the trial-by-trial prediction task).

In orders A and B, the measures of objective causal strength (*causal strength* and *predictiveness*) were presented before the measure of subjective feelings-of-success (*predictive success*). Between these two orders, the two measures of objective causal strength were counterbalanced. In orders C and D, participants were first presented with the measure of subjective feelings-of-success, followed by the two measures of objective causal strength (counterbalanced).

4.2 Results

4.2.1 Prediction accuracy

As in the prior experiments, participants' prediction accuracy was not the main focus of the analyses reported here, but was measured to ensure participants were properly learning about the causal relationships from the data. The trial-level prediction accuracy of participants (reported in Appendix B) showed that participants were increasingly accurate in predicting the state of the effect from the cause over time. Similar to the prior experiments, participants were learning more about the causal relationship with more trials of the prediction task.

A further reason for measuring participants' prediction accuracy was to compare their performance in the prediction task across conditions. The bonus points won by participants for each scenario are displayed in Figure 12.

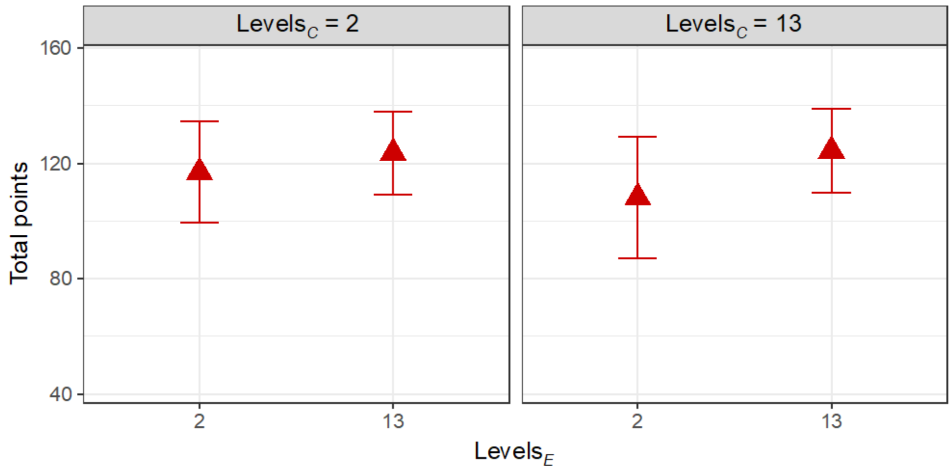

Figure 12. **Bonus points won by participants in each scenario by the granularity of the cause and effect in Experiment 3. Triangles represent condition means, and error bars represent standard deviations.**

To assess if there were differences in prediction accuracy at the scenario-level, a regression was run predicting the total bonus points won in each scenario from Levels_C and Levels_E, with by-subject random slopes for Levels_C and Levels_E to account for repeated measures. There was a small but significant effect of Levels_C ($B = -3.96$, 95% CI [-6.42, -1.50], $p = .002$, $R^2_{NSJ} = .013$, 95% CI [.001, .036]); participants scored slightly more points when the cause had fewer levels. There were significant differences in total points across conditions with different Levels_E ($B = 11.46$, 95% CI [8.62, 14.30], $p < .001$, $R^2_{NSJ} = .099$, 95% CI [.060, .146]); participants scored more points when the effect had more levels.

Similar to Experiments 1 and 2, the observed differences in bonus points won per scenario were in the intended direction. Because there were differences across conditions, when assessing the influence that the granularity of the cause and effect has on judgments, I performed analyses of both the raw judgments and also repeated the analyses after controlling for the bonus points. In addition, I controlled for the bonus points won within each scenario in the mediation analysis.

4.2.2 Testing the specificity criterion, the feelings-of-success heuristic, and the proportionality criterion

To recap, if participants were sensitive to the specificity criterion, their judgments would be sensitive to the granularity of the cause (Levels$_C$). If participants used the feelings-of-success heuristic, their judgments would be sensitive to the granularity of the effect (Levels$_E$). If participants were sensitive to the proportionality criterion, their judgments would be greater in proportional conditions compared to non-proportional conditions.

First, I ran regressions for each of the measures (*causal strength*, *predictiveness*, and *prediction success*), predicting each from Levels$_C$ and Levels$_E$ (treated as factors), and the grouping variable indicating proportionality. The regressions controlled for the objective causal strength of the datasets, and included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures. Next, I ran a second set of regressions that were identical to the first set, but controlled for the total bonus points won in each scenario.

The results for both sets of regressions are presented in Table 7. Figure 13 displays participants' judgments after controlling for predictors in each regression model other than Levels$_C$ and Levels$_E$. Figure 14 displays participants' judgments after controlling for all predictors other than the grouping variable indicating if a condition was proportional or non-proportional.

Table 7. ***Multivariate Regression Results for Experiment 3***

Predictor	Not controlling for bonus points			Controlling for bonus points		
	B	*p*	R^2_{NSJ}	*B*	*p*	R^2_{NSJ}
Measure: *Causal strength*						
$Levels_C$	-0.52 [-0.83, -0.22]	.001	.015 [.002, .039]	-0.19 [-0.33, -0.05]	.008	.006 [.000, .023]
$Levels_E$	-0.76 [-1.09, -0.44]	< .001	.031 [.010, .062]	-1.30 [-1.48, -1.11]	< .001	.197 [.147, .251]
Proportionality	0.29 [-0.02, 0.59]	.065	.004 [.000, .021]	0.06 [-0.08, 0.21]	.380	.001 [.000, .011]
Measure: *Predictiveness*						
$Levels_C$	-0.20 [-0.40, -0.01]	.044	.005 [.000, .022]	-0.04 [-0.22, 0.13]	.619	.000 [.000, .009]
$Levels_E$	-0.96 [-1.20, -0.72]	< .001	.102 [.062, .149]	-1.39 [-1.60, -1.18]	< .001	.200 [.150, .255]
Proportionality	0.20 [0.01, 0.40]	.043	.005 [.000, .022]	0.02 [-0.15, 0.19]	.837	.000 [.000, .008]
Measure: *Prediction success*						
$Levels_C$	-0.35 [-0.52, -0.18]	< .001	.016 [.003, .041]	-0.38 [-0.69, -0.08]	.013	.009 [.000, .028]
$Levels_E$	-0.88 [-1.09, -0.66]	< .001	.095 [.056, .140]	-1.13 [-1.47, -0.80]	< .001	.064 [.032, .104]
Proportionality	0.25 [0.08, 0.41]	.005	.008 [.000, .027]	0.13 [-0.17, 0.43]	.413	.001 [.000, .012]

Note. Confidence intervals represent 95% CIs.

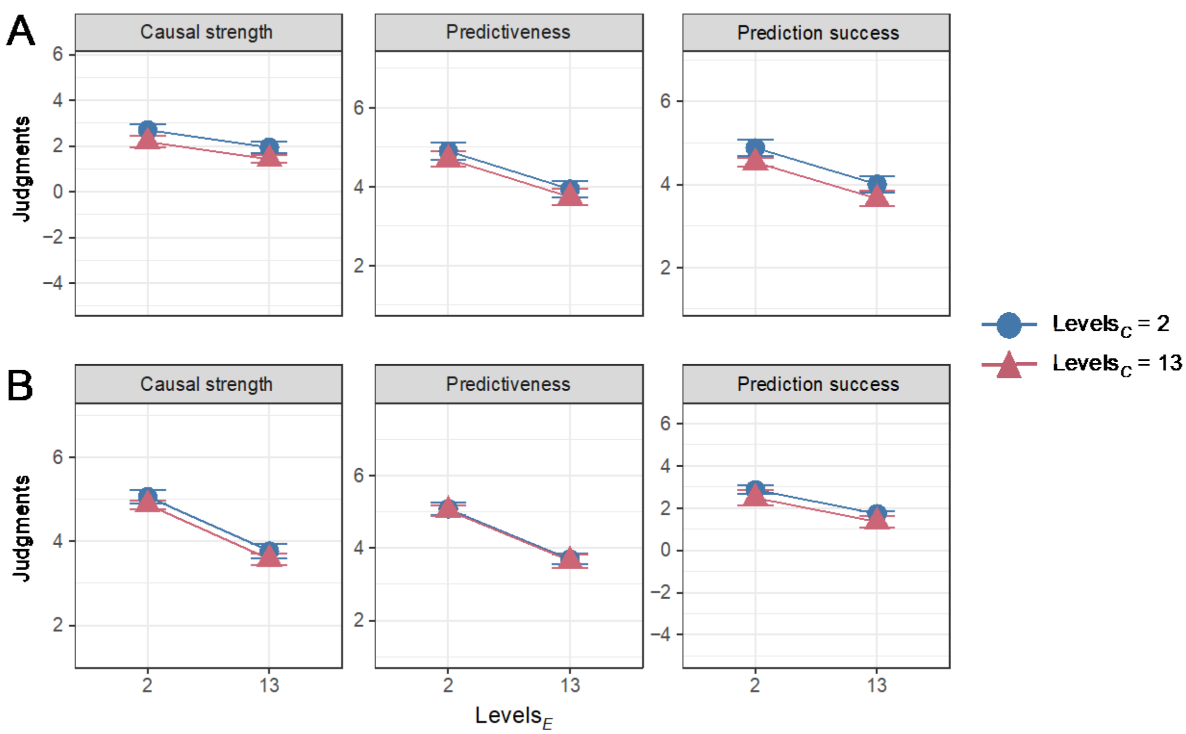

Figure 13**. Participants' judgments by $Levels_C$ and $Levels_E$ after controlling for other predictors in the two sets of regression models in Experiment 3. Averages are displayed over raw judgments; error bars indicate 95% CIs. . (A) Judgments without controlling for bonus points. (B) Judgments after controlling for bonus points.**

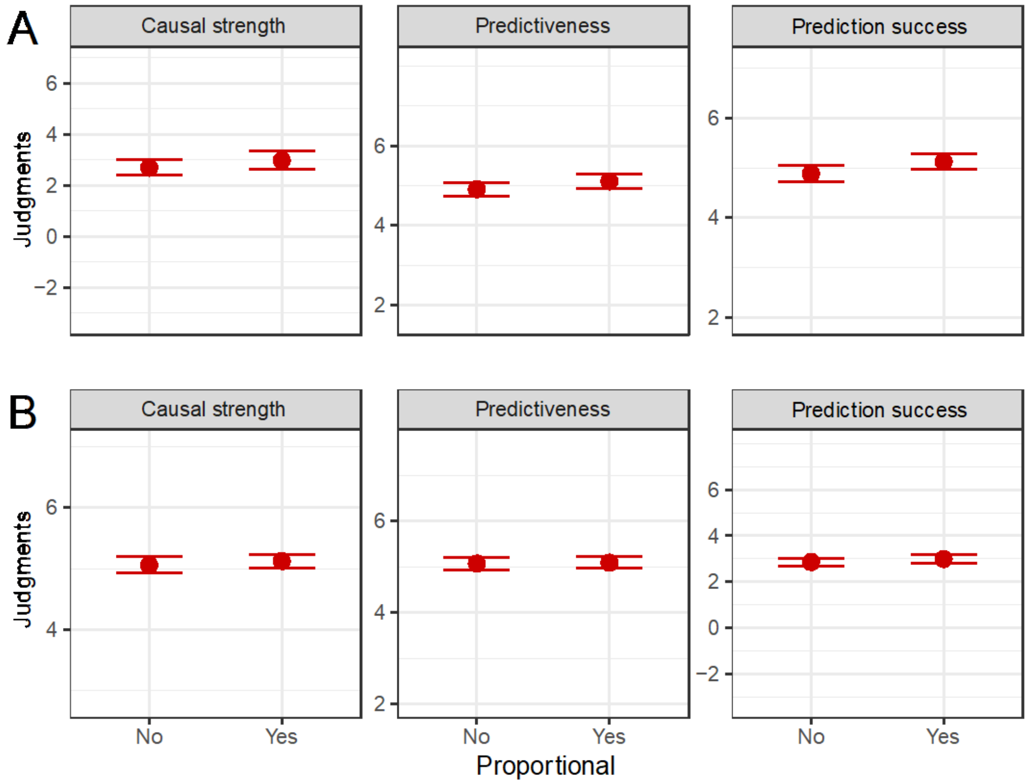

***Figure 14*. Participants' judgments for proportional vs. non-proportional conditions after controlling for other predictors in the two sets of regression models in Experiment 3. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments without controlling for bonus points. (B) Judgments after controlling for bonus points.**

Similar to both Experiments 1 and 2, the granularity of the effect was a significant predictor of judgments on all measures, both before and after controlling for the bonus points won per scenario. Participants made weaker judgments when the effect had more levels, as implied by the negative coefficients for Levels$_E$ in Table 7 and the decreasing slopes in Figure 13. This result was consistent with participants using the feelings-of-success heuristic to make judgments about the data.

Unlike in the prior experiments, in the regressions that did not control for bonus points, participants made weaker judgments on all measures when the cause had more levels. After controlling for bonus points, the influence of the cause's granularity on judgments of *predictiveness* became non-significant, while they remained significant for judgments of *causal strength* and *prediction success*. The influence of the cause's granularity was the opposite of what was predicted by the specificity criterion; participants gave stronger ratings for more coarse-grained causes. Despite these results diverging results from prior experiments, relative to the influence of the granularity of the effect, the effect sizes for Levels_C were very small (accounting for less than 1% of unique variance after controlling for prediction accuracy). In contrast, the effect sizes for Levels_E were substantially larger.

Before controlling for bonus points, participants made greater judgments of *prediction success*, and marginally greater judgments of *causal strength* and *predictiveness* in the proportional vs. non-proportional conditions. However, after controlling for bonus points, there were no differences between the conditions for all measures. Again, there was no clear pattern indicating that participants were sensitive to the proportionality criterion.

4.2.3 Mediation analysis testing subjective feelings-of-success as a mechanism

To test whether subjective feelings-of-success served as a mechanism for how the granularity of the effect influenced judgments of objective causal strength, I conducted a mediation analysis using the mediation package in R (Tingley, Yamamoto, Hirose, & Keele, 2014). This package allowed mediation analyses in repeated measures designs involving regressions with random slopes and intercepts. This analysis was conducted by fitting regression models between Levels_E and the mediator (*prediction success* judgments), and between Levels_E and the outcome variable of interest (either *causal strength* or *predictiveness* judgments) after controlling for the mediator. Each of these regressions also controlled for Levels$_C$, proportionality, the objective causal strength, and the bonus points won per scenario. Each regression also included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures.

First, I investigated whether participants' judgments of *prediction success* mediated the relationship between Levels$_E$ and judgments of *causal strength*. There was a significant average causal mediation effect, which meant that the entire pathway from Levels$_E$ through judgments of *prediction success* to judgments of *causal strength* was significant (ACME = -0.54, 95% CI [-0.74, -0.36], $p < .001$). There was also a significant average direct effect (ADE = -0.60, 95% CI [-0.97, -0.24], $p < .001$) from Levels$_E$ to judgments of *causal strength*.

Second, I analyzed whether participants' judgments of *prediction success* mediated the relationship between Levels$_E$ and judgments of *predictiveness*. Again, there was a significant average causal mediation effect (ACME = -0.93, 95% CI [-1.10, -0.78], $p < .001$), as well as a significant average direct effect (ADE = -0.44, 95% CI [-0.63, -0.26], $p < .001$).

The results for the mediation analyses (including all order conditions) are displayed in Figure 15 to help visualize the effects in terms of the proposed mechanistic model. The negative coefficients indicate that for more fine-grained effects, participants made weaker judgments of *causal strength* and *predictiveness.*

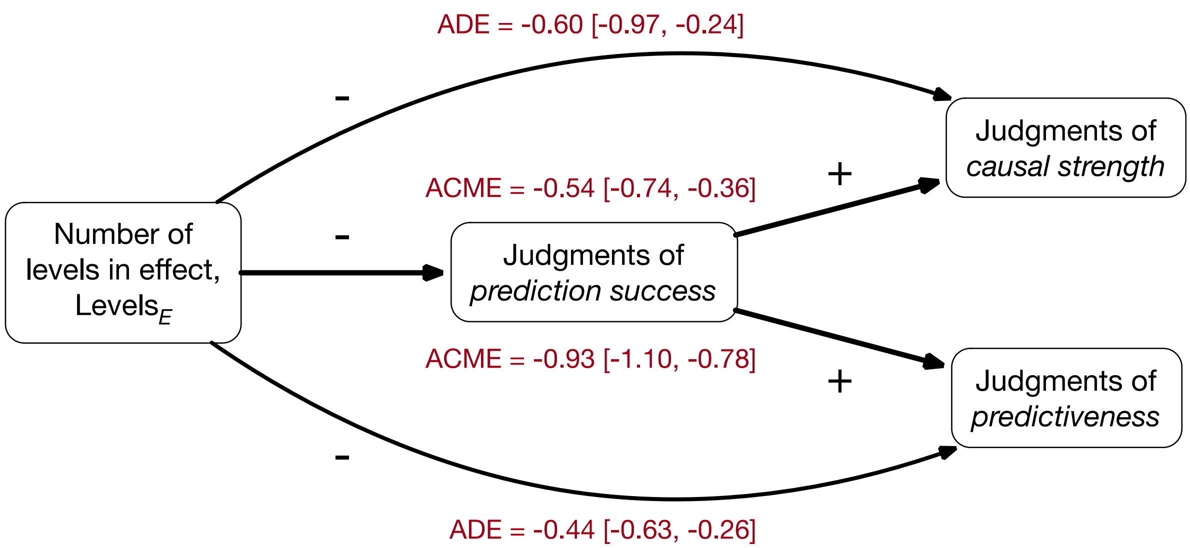

***Figure 15*. Mediation analysis results for the relationship between Levels$_E$, judgments of *prediction success*, and judgments of causal *strength* and *predictiveness* in Experiment 3. All effects are significant at the *p* < .001 level. ACME = Average Causal Mediation Effect. ADE = Average Direct Effect.**

To test if the mediation effects reported above were due to any order effects, I repeated the analyses above for each order condition (see Table 6). The results of these analyses are presented in Table 8.

Table 8. ***Results for Mediation Analysis Results by Order Condition in Experiment 3***

Measure		Order of measures	ACME	*p*	ADE	*p*
Causal strength	A	*Causal strength, predictiveness, prediction success*	-0.57 [-0.97, -0.23]	< .001	-0.30 [-1.15, 0.51]	.494
	B	*Predictiveness, causal strength, prediction success*	-0.80 [-1.23, -0.40]	< .001	-0.18 [-0.92, 0.60]	.634
	C	*Prediction success, causal strength, predictiveness*	-0.59 [-0.99, -0.20]	.004	-0.50 [-1.28, 0.26]	.214
	D	*Prediction success, predictiveness, causal strength*	-0.33 [-0.70, -0.04]	.032	-1.31 [-2.03, -0.58]	.002
Predictiveness	A	*Causal strength, predictiveness, prediction success*	-1.02 [-1.37, -0.66]	< .001	-0.27 [-0.59, 0.06]	.120
	B	*Predictiveness, causal strength, prediction success*	-0.83 [-1.21, -0.49]	< .001	-0.29 [-0.70, 0.14]	.190
	C	*Prediction success, causal strength, predictiveness*	-1.06 [-1.39, -0.75]	< .001	-0.58 [-0.99, -0.16]	.004
	D	*Prediction success, predictiveness, causal strength*	-0.82 [-1.17, -0.50]	< .001	-0.59 [-0.94, -0.22]	.002

Note. ACME = Average Causal Mediation Effect. ADE = Average Direct Effect. Confidence intervals represent 95% CIs.

Regardless of the order condition, the average causal mediation effect on both judgments of *causal strength* and *predictiveness* were significant, or at least marginally significant for judgments of *causal strength* in order condition D.[8] These results indicated that the causal mediation effects were robust and not due to any particular order in which participants made judgments.

The fact that the causal mediation effects were present even when participants made judgments of *prediction success* last (orders A and B) suggested that participants formed subjective feelings-of-success prior to making a judgment about them, and that these feelings influenced the judgments of objective strength they made first. The usual strategy in mediation analyses is to measure the variables in their causal order, i.e. measuring judgments of *prediction success* before judgments of *causal strength* and *predictiveness*. Orders A and B were atypical in that the proposed mediator (*prediction success*) was measured last, but were necessary to demonstrate that subjective feelings-or-success were not simply a result of participants being prompted to make judgments of *prediction success*.

In sum, the mediation results show that the relationship between the granularity of the effect and participants' estimates of the objective strength in the data was at least partially mediated by participants' levels of subjective feelings-of-success. In contrast, while the average direct effects were significant when collapsing across all order conditions, the effects were not robust to the order of measures that participants made judgments for.

[8] The separate mediation analyses for each order involved only 40 participants per condition, so it is unsurprising that there was some variation across the different orders.

4.3 Discussion

In Experiment 3, both the granularity of the cause and effect influenced participants' judgments on all measures. However, as before, the dominant influence on participants' judgments was from the granularity of the effect; participants made stronger judgments in conditions with more coarse-grained effects.

The granularity of the effect has a bearing on how easy or difficult the prediction task feels to participants; when making predictions of a coarse-grained effect it feels more difficult to make an accurate prediction, leading to lower subjective feelings-of-success. I proposed that the subjective feelings-of-success (captured by the *prediction success* measure) might serve as a mechanism for the relationship between the granularity of the effect and judgments of objective strength. Results from the mediation analyses revealed that judgments of *prediction success* indeed significantly mediated the relationships between the granularity of the effect on one hand, and judgments of *causal strength* and *predictiveness* on the other.

The mediation effects were present regardless of the order of measures. Beyond demonstrating the robustness of the effects, the findings shed further light on the nature of subjective feelings-of-success as a mechanism. The fact that the causal mediation effects were present even when participants made judgments of *prediction success* last suggested that participants formed subjective feelings-of-success prior to making a judgment about them, and that these feelings influenced their judgments of objective strength. In other words, subjective feelings-of-success arise spontaneously when participants engaged in the trial-by-trial prediction task, and not simply as a result of responding on the *prediction success* measure.

Identifying subjective feelings-of-success as a mechanism helps to explain why judgments of objective strength varied with the manipulation of the granularity of the effect, despite the actual strength between the cause and effect being fixed in the datasets participants experienced.

The significant direct effects in some order conditions suggest further mechanisms (other than subjective feelings-of-success that they got from accurate predictions) that mediate participants' judgments of the objective strength. One plausible mechanism is that when participants obtained higher bonuses, they felt that they were better able to predict the effect and that the causal relation was stronger. In fact, performing the same mediation analyses but using the bonus points won per-scenario instead of judgments of *prediction success* as a mediator revealed significant causal mediation effects on both *causal strength* and *predictiveness* (p's $<$ .001). Both mechanisms (subjective feelings-of-success and bonus points) are consistent with the notion that participants made judgments of a relatively inaccessible quantity (objective strength) by substituting a more easily accessible one.

There are a couple of reasons why participants may substitute subjective feelings-of-success for judgments of objective strength. First, subjective feelings-of-success may be accessible because they are tied to the ongoing experience of the trial-by-trial prediction task. On each trial, participants have a sense of their prediction accuracy (via subjective feelings-of-success and the bonus points) which can be updated on each trial. In contrast, it is more difficult for participants to assess the objective strength (either in terms of *causal strength* or *predictiveness*), as it requires attending to the actual states of the cause and effect across all trials while ignoring the predictions they made. It is also possible that participants do not fully attend to or remember the actual states of the effect, focusing instead on the states they predicted.

A second reason participants may substitute subjective feelings-of-success for judgments of objective strength is that it may be useful in many situations to do so. Subjective feelings-of-success may serve as a sensible cue of objective strength because stronger (and better) causes should intuitively allow better success in the prediction task. Thus, if participants experienced greater subjective feelings-of-success, they may have reasoned that it was due to greater strength in the causal relationship.

In sum, participants' subjective feelings-of-success that arose due to the prediction task served as a substitute for judgments of objective strength. While this substitution may prove useful in many contexts due to the intuitive relation between ease of prediction and the strength of a cause, it can lead to errors when subjective feelings-of-success are artificially influenced by the granularity of the effect.

5.0 Experiment 4: How the act of prediction moderates the influence of granularity

The novel finding in Experiment 3 was that subjective feelings-of-success mediated the relationship between the granularity of the effect and judgments of the objective strength of a causal relationship. This finding supported the theory that participants assessed objective strength (measured via judgments *causal strength* and *predictiveness*) by substituting a more accessible quantity like subjective feelings-of-success (measured via judgments of *prediction success*). Subjective feelings-of-success may be more accessible because participants derive it from their experience of using the cause to predict the effect throughout the scenario.

In Experiment 4, I investigated the source of participants' subjective feelings-of-success. The findings from Experiments 1-3 are consistent with the hypothesis that participants' judgments are driven by their subjective feelings-of-success; when the effect is coarse-grained, participants might feel they are performing well in the prediction task because there is a higher chance of predicting the correct state of the effect by chance. Thus, the influence of granularity likely arises because participants are engaging in the act of prediction, using variables that have differing numbers of levels.

The present experiment investigates if making explicit predictions is necessary for participants to experience subjective feelings-of-success. In Experiments 1–3, the trial-by-trial prediction task forced participants to make predictions of the effect. However, research across various domains of cognition reveals that people automatically engage in prediction – when given information about some variable, people make unprompted predictions about related quantities (e.g., Nisbett & Borgida, 1975; Sebanz & Knoblich, 2009; Tversky & Kahneman, 1983). It is possible that the tendency people have to formulate predictions will lead to them making predictions even when not explicitly forced to. If so, participants' causal judgments could be influenced by the granularity of the effect even when participants do not explicitly predict the state of the effect from the cause.

To understand the role of the trial-by-trial prediction task in producing participants' subjective feelings-of-success, I tested whether or not the influence of the effect's granularity would influence participants' judgments of objective strength across three learning paradigms. In one paradigm (used in Experiments 1-3), participants make an explicit prediction of the effect. In a second paradigm, participants do not make an explicit prediction, but likely still make predictions after seeing the effect. In the third paradigm, participants see the cause and effect simultaneously, so they cannot make a prediction. If the granularity of the effect influences participants' judgments mainly through the explicit process of prediction, then the effect would only hold for the first condition. Alternatively, it might be strongest when making explicit predictions, weaker when the task allows for but does not require predictions, and weakest when the task prohibits predictions. Lastly, if the influence of granularity of the effect is due to some other process, then it could hold for all three conditions.

5.1 Method

5.1.1 Participants

240 new participants were recruited on Mturk using the same criteria as in the prior experiments. Each participant was paid $0.90; there were no accuracy bonuses. The experiment lasted between six to eight minutes.

5.1.2 Design and procedure

Similar to Experiments 2 and 3, the present experiment utilized a within-subject manipulation of the number of levels in the cause (Levels$_C$) and effect (Levels$_E$); each variable had either two or 13 levels (resulting in four within-subject conditions).

In the present experiment, the key manipulation (between-subjects) was the learning paradigm by which participants experienced each dataset; prediction task vs. simultaneous observation vs. delayed observation. In the prediction task condition, participants learned about the cause and effect by engaging in the trial-by-trial prediction task, without receiving bonus points (identical to the no-bonus condition in Experiment 2). Participants in the prediction task condition were not given bonus points for consistency with the observation conditions, in which participants did not make predictions, and therefore received no bonus points.

In the delayed observation condition, on each trial, the state of the cause was immediately revealed, followed by the effect after one second (Figure 16). This allowed time for participants to implicitly predict the effect, but did not require them to do so. After viewing each observation, participants clicked on a button to advance to the next trial. In many ways, this condition most closely aligns with real-world learning, in that causes occur before effects (and usually with some delay between them), allowing people to predict the effect. This condition is also somewhat analogous to real-world situations in which a person makes an action, which is then followed by an outcome, with the difference being that participants do not make a choice in this condition.

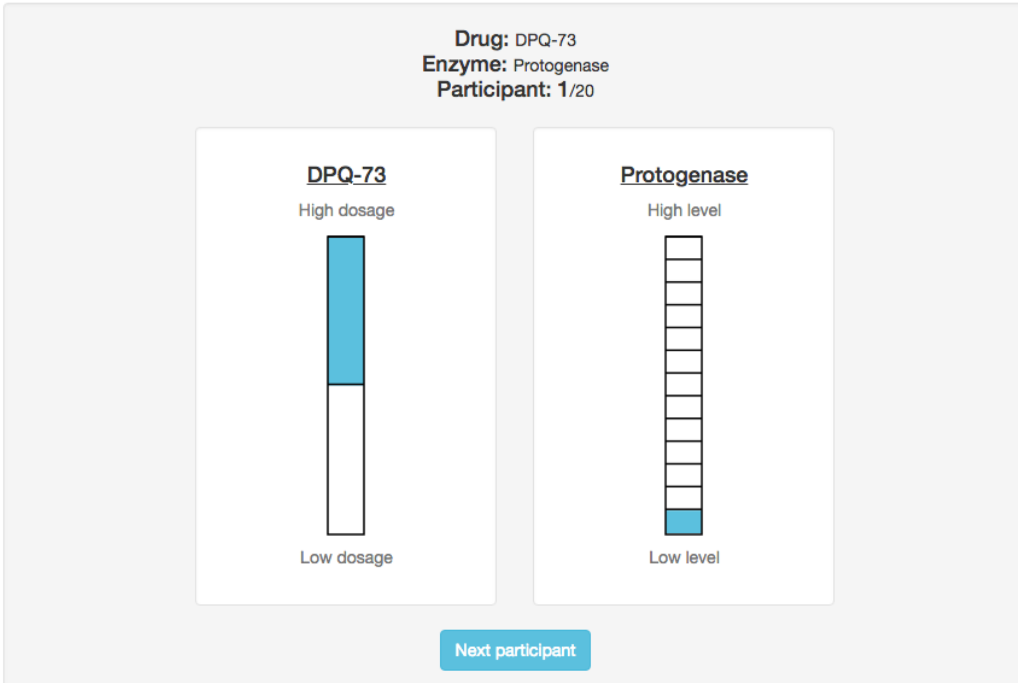

Figure 16. **Presentation of stimuli in the observation conditions in Experiment 4. The actual states of the cause and effect are revealed on each trial without participants making any prediction. In the delayed observation condition, there was a one-second delay between when the state of the cause was revealed and the state of the effect was revealed.**

In the simultaneous observation condition, on each trial, the states of the cause and effect were revealed simultaneously. After viewing each observation, participants clicked on a button to advance to the next trial. In this condition, it is impossible to make a prediction of the effect based on the cause. If the influence of the granularity of the effect still arises in this condition, then it must not be due exclusively to engaging in the act of prediction. In both the observation conditions, the button to advance to the next trial was disabled for two seconds after pressing it.

Unlike prior experiments, participants in Experiment 4 only made judgments using either the *causal strength* or *predictiveness* measures. The *prediction success* measure was omitted because it was irrelevant in the observation conditions, in which participants did not make any predictions. For consistency, it was also omitted from the prediction task condition as well. Similar to Experiments 1 and 2, the measures were presented between-subjects. In combination with the learning format, there were six between-subject conditions, with 40 participants assigned to each condition.

5.2 Results

5.2.1 Testing the specificity criterion, the feelings-of-success heuristic, and the proportionality criterion

The main purpose of the analyses was to determine if finer granularity in the effect leads to weaker judgments of objective strength would be moderated by the learning paradigm. If subjective feelings-of-success arise as a result of the prediction task, I expected that similar to prior experiments, finer granularity in the effect would lead to weaker judgments in the prediction task condition. I predicted that this influence would be attenuated in the simultaneous observation condition because participants would not be able to make any predictions due to the actual states of the cause and effect being revealed simultaneously. In the delayed observation condition, the delay potentially gave participants time to make a spontaneous prediction of the effect after the cause was revealed even though the learning task did not elicit a prediction. If participants made predictions of the effect spontaneously after viewing the cause, the influence of the effect's granularity in this condition would be similar to the prediction task condition. If participants did not make spontaneous predictions, the influence of the effect's granularity would be similar to the simultaneous observation condition.

For each of the three learning paradigm conditions, I ran regressions for each of the measures used in this experiment (*causal strength* and *predictiveness*) predicting each from Levels_C and Levels_E, and a grouping variable indicating whether a particular condition had proportional vs. non-proportional causes and effects. The regressions controlled for the objective causal strength of the datasets, and included by-subject random slopes for Levels_C and Levels_E to account for repeated measures.

The results of these regressions are presented in Table 9. Figure 17 displays the influence of Levels_C and Levels_E on participants' judgments. Figure 18 displays the influence of proportionality on participants' judgments.

Table 9. ***Multivariate Regression Results for Experiment 4***

Predictor	Prediction task			Delayed observation			Simultaneous observation		
	B	*p*	R^2_{NSJ}	*B*	*p*	R^2_{NSJ}	*B*	*p*	R^2_{NSJ}
Measure: *Causal strength*									
Levels_C	-0.37 [-0.95, 0.22]	.222	.006 [.000, .054]	-0.89 [-1.71, -0.06]	.042	.026 [.000, .094]	-0.70 [-1.41, 0.02]	.064	.025 [.000, .093]
Levels_E	-0.79 [-1.37, -0.21]	.008	.029 [.001, .100]	-0.14 [-0.88, 0.60]	.716	.001 [.000, .034]	-0.21 [-0.81, 0.38]	.480	.002 [.000, .041]
Proportionality	0.14 [-0.44, 0.72]	.642	.001 [.000, .035]	-0.04 [-0.73, 0.65]	.915	.000 [.000, .031]	0.83 [0.25, 1.41]	.007	.035 [.001, .109]
Measure: *Predictiveness*									
Levels_C	-0.31 [-0.69, 0.06]	.102	.014 [.000, .071]	-0.41 [-0.87, 0.05]	.090	.017 [.000, .079]	-0.32 [-0.72, 0.08]	.122	.011 [.000, .066]
Levels_E	-1.96 [-2.41, -1.51]	< .001	.353 [.247, .461]	-0.55 [-1.00, -0.10]	.019	.032 [.001, .104]	-0.40 [-0.83, 0.03]	.077	.018 [.000, .079]
Proportionality	-0.13 [-0.49, 0.23]	.469	.003 [.000, .042]	0.21 [-0.23, 0.64]	.349	.005 [.000, .049]	0.26 [-0.10, 0.62]	.164	.007 [.000, .057]

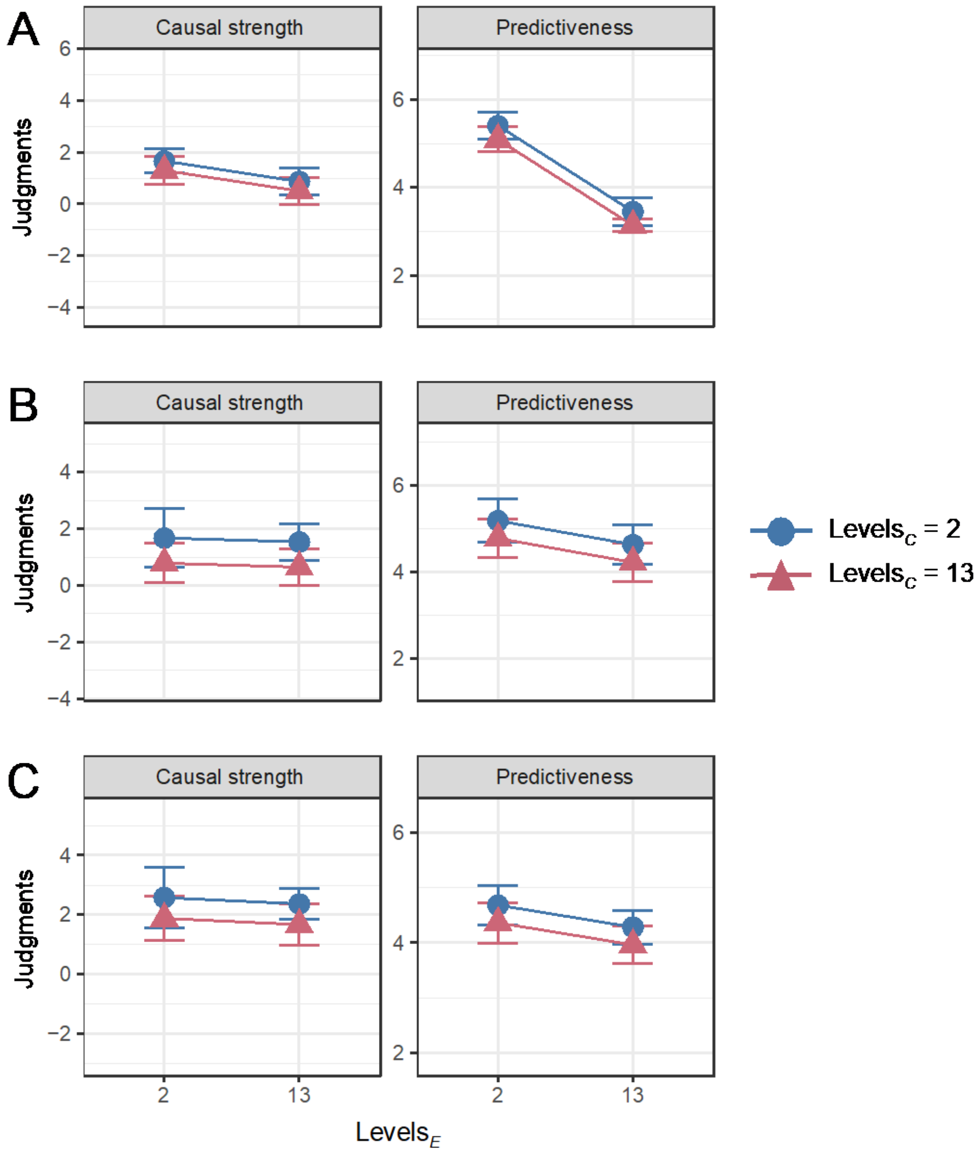

Figure 17. **Participants' judgments by Levels$_C$ and Levels$_E$ after controlling for other predictors in the regression models in Experiment 4. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments from prediction task condition. (B) Judgments from delayed observation condition. (C) Judgments from simultaneous observation condition.**

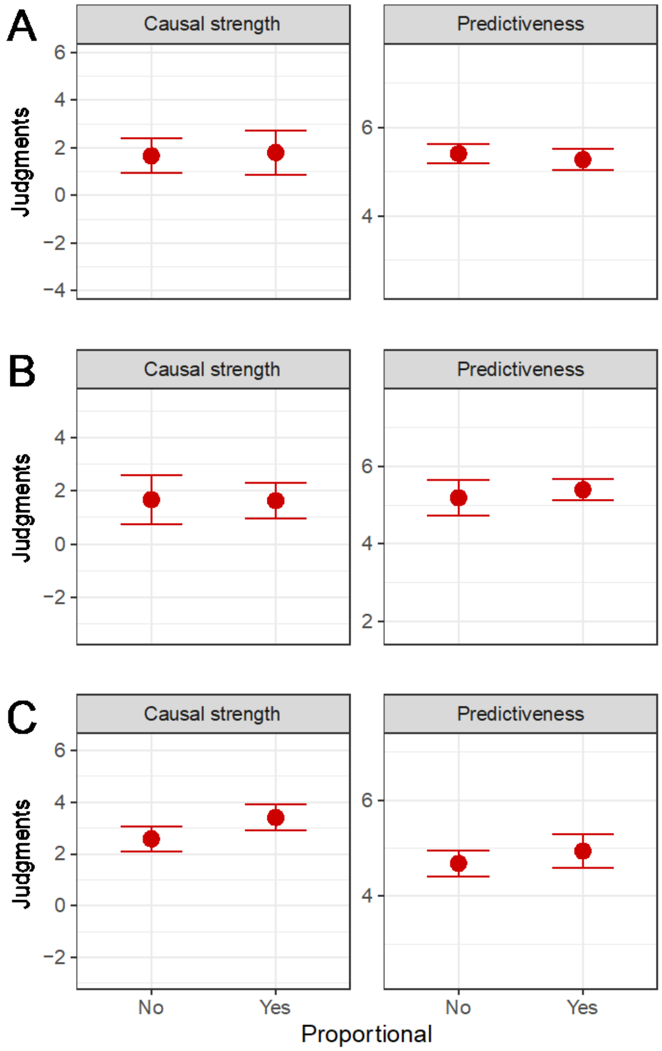

***Figure 18*. Participants' judgments for proportional vs. non-proportional conditions after controlling for other predictors in the regression models in Experiment 4. Averages are displayed over raw judgments; error bars indicate 95% CIs. (A) Judgments from prediction task condition. (B) Judgments from delayed observation condition. (C) Judgments from simultaneous observation condition.**

5.2.1.1 The influence of Levels$_C$

Similar to Experiments 1 and 2, the granularity of the cause (Levels$_C$) was not a significant predictor of either objective strength measure in the prediction task condition (Table 9). In the delayed observation condition, Levels$_C$ was a marginally significant predictor of judgments of *causal strength* but not of judgments of *predictiveness.* In the simultaneous observation condition, Levels$_C$ was a marginally significant predictor of judgments of *causal strength* but not of *predictiveness.* For both observation conditions, participants gave slightly stronger ratings of *causal strength* for more coarse-grained causes, which was the opposite of what was predicted by the specificity criterion.

To test if the learning paradigm moderated the influence of Levels$_C$, I ran separate regressions using data from all learning paradigm conditions, predicting each measure of objective strength from Levels$_C$, Levels$_E$, proportionality, the learning paradigm condition, and the two-way interaction between Levels$_C$ and the learning paradigm condition factor. The regressions included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures, and controlled for the objective causal strength of the datasets. The learning paradigm did not moderate the influence of Levels$_C$ for judgments of *causal strength* (p = .561) or *predictiveness* (p = .898).

In sum, taken together with the varied findings regarding the influence of Levels$_C$ across the prior experiments, there are no consistent patterns showing people are sensitive to the specificity criterion.

5.2.1.2 The influence of Levels$_E$

As expected, in the prediction task condition, participants made stronger judgments of *causal strength* and *predictiveness* when the effect had fewer levels (Table 9). In the delayed observation condition, the influence of Levels$_E$ was significant for judgments of *predictiveness* but not judgments of *causal strength*. In the simultaneous observation condition, the influence of Levels$_E$ was marginally significant for judgments of *predictiveness*, but not significant for judgments of *causal strength*.

To test whether the learning paradigm moderated the influence of Levels$_E$, I ran separate regressions using data from all learning paradigm conditions, predicting each measure of objective strength from Levels$_C$, Levels$_E$, proportionality, the learning paradigm condition, and the two-way interaction between Levels$_E$ and the learning paradigm condition. The regressions included by-subject random slopes for Levels$_C$ and Levels$_E$ to account for repeated measures, and controlled for the objective causal strength of the datasets.

The learning paradigm did not moderate the influence of Levels$_E$ for judgments of *causal strength* (p = .283). However, for judgments of *predictiveness*, the influence of Levels$_E$ was stronger in the prediction task condition compared to the delayed observation condition (B = 1.35, 95% CI [0.74, 1.96], p < .001, R^2_{NSJ} = .034, 95% CI [.010, .073]) and the simultaneous observation condition (B = 1.55, 95% CI [0.94, 2.16], p < .001, R^2_{NSJ} = .045, 95% CI [.016, .087]). There was no difference in the influence of Levels$_E$ between the delayed and simultaneous observation conditions (p = .518).

In sum, when making judgments of *causal strength*, there was only an influence of Levels$_E$ when participants made explicit predictions. I had predicted that the influence of Levels$_E$ would be present, although weaker in the delayed observation condition. When making judgments of *predictiveness*, the pattern of results closely aligned with the hypotheses: the influence of Levels$_E$ was strongest for the prediction condition, still significant although weaker for the delayed observation condition, and not significant for the simultaneous observation condition.

The biggest inconsistency with the hypotheses was that there was a significant effect of Levels$_E$ for the delayed condition for *predictiveness* but not for *causal strength*. One possible explanation for this is that the influence of Levels$_E$ applies when participants are focusing on how well the causal relationship enables prediction – i.e. when making judgments of *predictiveness*. In contrast, when making judgments of *causal strength*, participants may be thinking of the strength of the relationship independent of its predictive utility.

5.2.1.3 The influence of proportionality

There were no differences in judgments of *predictiveness* between proportional and non-proportional conditions for any of the learning paradigm conditions (Table 9). Participants made stronger judgments of *causal strength* in the proportional condition in the simultaneous observation condition. There were no differences in judgments of *causal strength* for the prediction task or delayed observation conditions.

The findings regarding proportionality in the prediction task condition were similar to Experiment 3, which also found marginal effects of proportionality for judgments of *causal strength*. However, across all experiments, the effect of proportionality is not reliable. Further, it is not clear what is driving these differences, especially across experiments. Given the inconsistency of findings, and the fact that the observation conditions in the present experiment are completely different from prior experiments, I did not perform analyses to investigate if the learning task moderated differences in judgments between the proportional vs. non-proportional conditions.

5.3 Discussion

In Experiment 4, participants' judgments of *causal strength* and *predictiveness* were influenced by the granularity of the effect when they made predictions while learning about the causal relationship. When participants learned about the causal relationship via observation with a delay between the cause and effect, their judgments of *predictiveness* were also influenced by the granularity of the effect. This result suggested that participants were making implicit predictions, despite the task not requiring them to do so. However, when learning about the causal relationship via simultaneous observation of the cause and effect, the granularity of the effect on judgments of *predictiveness* was marginally significant. Despite having no opportunity to make a prediction, it is possible that participants first attended to the state of the cause and implicitly began to form a prediction before attending to the state of the effect. The results imply that subjective feelings-of-success, the mechanism by which the effect's granularity influences judgments of objective strength, arise mainly when participants engage in the act of prediction.

In the prior experiments, participants were made to learn about a causal relationship via a prediction task for several reasons. First, this task was used for consistency with prior causal learning research involving trial-by-trial presentations of data (e.g., Derringer & Rottman, 2018; Liljeholm, 2015; Spellman, 1996). Second, I expected that making predictions on each trial would lead to participants being more engaged and attentive to the data. Third, prediction is a common goal and function of causal knowledge in everyday contexts (e.g., Grzymala-Busse, 2011; Johnson & Keil, 2014; Zellner, 1988).

Learning while engaging in the prediction task could be viewed as a more *active* form of causal learning because participants are given the opportunity to make inferences about the data. In past research on causal learning, a distinction has often been drawn between learning via passive observation of summary data (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Jenkins & Ward, 1965; Vasilyeva, Blanchard, & Lombrozo, 2016) and learning via attending to sequential data that arises from interventions on a causal system (e.g., Hagmayer & Meder, 2013; Hagmayer, Meder, Osman, Mangold, & Lagnado, 2010; Hagmayer & Sloman, 2009; Robinson, Sloman, Hagmayer, & Hertzog, 2010; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005). The latter can be considered an active form of learning, as it requires learners to formulate and test hypotheses about the causal relationships in order to learn causal structures (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Lagnado & Sloman, 2004).

The act of prediction may be inherent to some extent in all forms of learning that are active. The present prediction task differs from the aforementioned studies in that the goal is not to learn causal structures, but rather to predict the outcome of a variable. However, both require learners to make predictions about the outcome of some variable given information about related variables. Future research should investigate the effect of granularity in other tasks involving forms of active causal learning beyond the prediction task used in the present experiments.

In sum, the present experiment, by comparing the influence of the effect's granularity across different learning paradigms, demonstrated that the influence of the granularity of variables on causal learning depends on the type of learning task participants are engaged in.

6.0 General discussion

The present research investigated the influence of varying levels of granularities of causes and effects on causal learning and reasoning. In order to navigate complex environments, people have to learn about causal relationships between variables to predict, explain, and control outcomes (Sloman, 2005; Sloman & Lagnado, 2015; Waldmann & Hagmayer, 2014). In the real world, some variables are more fine-grained (i.e. have more levels) and some are more coarse-grained (i.e. have fewer levels). Across all experiments reported here, participants were presented with data of cause-effect pairs with a fixed causal strength, while the granularities of the cause and effect were manipulated. Participants learned the strength of the causal relationship while predicting the states of the effect from states of the cause.

6.1 Summary of results

The present experiments consistently found that coarser-grained effects led to stronger judgments of the objective strength of the relationship between the cause and effect, and also stronger judgments of subjective feelings-of-success (FOS) based on how accurate participants were in predicting the effect. This finding is consistent with the newly-proposed *feelings-of-success heuristic*. However, there was inconsistent evidence about whether these dependent variables were influenced by the granularity of the cause or whether the granularity of the cause matched the granularity of the effect. These lack of consistent findings speak against the specificity criterion

(Franklin-Hall, 2016; Griffiths, et al., 2015; Weslake, 2010) and the proportionality criterion (McGrath, 1998; Woodward, 2010; Yablo, 1992), respectively.

In Experiment 1, I found that more fine-grained effects led to participants making weaker judgments of both subjective feelings of how accurate participants were in predicting the effect (the *prediction success* measure) and the objective strength of the relationship between the cause and effect, as measured by the *causal strength* and *predictiveness* measures. In Experiment 2, this finding held regardless of whether participants received or did not receive a bonus tied to the accuracy of their predictions.

I hypothesized that the pattern of results described above could occur for the following reason. When participants made predictions of finer-grained effects, they may have had lower subjective FOS; they likely perceived the prediction task as being more difficult and their prediction accuracy to be lower because they were less likely to predict the exact state of the effect. I further hypothesized that participants used their subjective FOS for judging the objective strength of the cause-effect relationship; this proposed substitution is the feelings-of-success heuristic.

I tested this hypothesis in Experiment 3 with a mediation analysis. Unlike in Experiments 1 and 2, in Experiment 3 participants made judgments using all three measures. Experiment 3 revealed that participants' judgments of *prediction success* indeed mediated the relationship between the granularity of the effect and judgments of *causal strength* and *predictiveness*.

The fact that participants' judgments of *causal strength* and *predictiveness* were influenced by the granularity of the effect in Experiments 1 and 2 suggests that participants may have experienced subjective FOS based on their performance in the trial-be-trial prediction task, which was in turn substituted for their estimates of the objective strength of the causal relationship, even though participants were not explicitly asked to make judgments of *prediction success*. The mediation analysis in Experiment 3 supported this interpretation.

The substitution of subjective FOS for estimates of the objective causal strength implied by the feelings-of-success heuristic is consistent with research showing people substitute subjective mental states for estimates of less accessible quantities (Hertzog et al., 2003; Oppenheimer, 2008). Subjective FOS may have been used as a cue to objective strength because the strength of a cause-effect relationship should generally be highly related to prediction accuracy. When a causal relationship is very strong, learners should be able to very accurately predict the state of the effect after viewing the cause. Conversely, when a causal relationship is very weak, learners should be less accurate in making predictions of the effect. Thus, from a rational perspective, it is sensible for learners to use the feelings-of-success heuristic. However, as the present experiments demonstrate, the granularity of the effect influences learners' subjective FOS independently of the objective causal strength; when the effect is very fine-grained, learners tend to judge themselves as having lower *prediction success*, which can explain their lower estimates of *causal strength* and *predictiveness*, despite the objective causal strength being held constant. In sum, the heuristic does not always lead to accurate inferences.

In Experiment 4, I investigated if the influence of the effect's granularity depended on participants engaging in the act of prediction. When participants were not required to make predictions explicitly, but observed the cause followed by the effect after a delay, the influence of the effect's granularity was weaker but still significant on judgments of *predictiveness*. The findings from Experiment 4 suggested that even when not explicitly required to make a prediction, participants were making implicit predictions of the state of the effect upon observing the state of the cause, which gave rise to subjective FOS depending on the effect's granularity.

6.2 The feelings-of-success heuristic

6.2.1 The pervasiveness of the feelings-of-success heuristic

I have proposed that when learning about a cause and effect while making predictions, the granularity of the effect influences learners' subjective FOS, which in turn influences their judgments of the objective causal strength. The way learners use subjective FOS as a heuristic for making causal judgments is similar to how people substitute more accessible subjective or affective mental states for more objective judgments (e.g., Forgas, 1995; Greifeneder, Bless, & Pham, 2011). As previously mentioned, people substitute feelings related to the ease of learning for judgments of how well they have learned something (Hertzog et al., 2003; Koriat, 2008), and perceived ease of recall for judgments of objective statistics related to a target (Goldstein & Gigerenzer, 2002; Hertwig et al., 2008; Kahneman & Tversky, 1973; Pachur & Hertwig, 2006).

The findings of Experiment 4 suggest that the granularity of the effect influences causal judgments largely because prediction is a pervasive cognitive act that occurs implicitly and spontaneously (e.g., Knoblich & Flach, 2001; Sebanz & Knoblich, 2009; Wolpert, Doya, & Kawato, 2003), giving rise to subjective FOS. In other words, the feelings-of-success heuristic may be a general heuristic applicable in learning tasks that require learners to make predictions and inferences different from the type of predictions involved in the present experiments.

While the granularity of the effect influences subjective FOS in the present experiments, there may be other factors that influence subjective FOS, which in turn would influence learners' judgments about objective causal strength. For example, increasing time pressure on learners as they make predictions may make the act of prediction feel more difficult and consequently diminish subjective FOS, leading to weaker causal judgments if learners are indeed using the feelings-of-success heuristic (see Finucane, Alhakami, Slovic, & Johnson, 2000; Rieskamp & Hoffrage, 2008; Suri & Monroe, 2004). Subjective FOS may also be diminished by making tasks more effortful if learners feel that they are performing worse due to increased demands of the task (Koriat & Ma'ayan, 2005; Koriat & Nussinson, 2009). In sum, if learners use the feelings-of-success heuristic more generally, numerous factors that influence subjective FOS could consequently influence judgments of objective quantities.

6.2.2 Subjective feelings-of-success in prior research

Subjective FOS may also have influenced the results of prior studies on causal learning in which participants were presented with trial-by-trial data of binary causes and binary effects. This may have occurred even when participants did not make explicit predictions, but subjective FOS likely had a larger effect when participants had to make explicit predictions concerning the state of the effect based on the state of the cause, as they are required to in many studies (e.g., Derringer & Rottman, 2018; Liljeholm, 2015; Spellman, 1996). If participants happened to make more correct predictions, they would have experienced greater subjective FOS, which would have led to stronger judgments of objective causal strength. In such cases, and in general, using the feelings-of-success heuristic by substituting subjective FOS for judgments of causal strength may be sensible, because stronger causal relationships should enable more accurate predictions.

The present experiments demonstrate that subjective FOS can be disentangled from objective causal strength by manipulating the granularity of the effect. Future research should investigate other factors that may influence subjective FOS independently from objective causal strength in order to identify the boundary conditions for the feelings-of-success heuristic being an effective way to accurately estimate causal strength.

Beyond research that specifically investigates how people assess the strength of a cause-effect relationship, subjective FOS may also influence judgments about covariation and associative strength between cues. Most models of associative and reinforcement learning posit that learners learn about relationships via making predictions about outcomes and receiving feedback on their prediction accuracy, which allows for the possibility that subjective FOS (arising from the act of making predictions) may influence estimates of associative strength (Klopf, 1988; Rescorla & Wagner, 1972; Rumelhart, Hinton, & Williams, 1986; Stone, 1986; Sutton, 1988; Sutton & Barto, 1981, 1987; also see Sutton & Barto, 1998). More generally, models of probability learning assume learners are making predictions about outcomes as they learn about probabilities (e.g., Biele, Erev, & Ert, 2009; Lejarraga, Dutt, & Gonzalez, 2010), while models of category learning assume learners are making predictions about features and properties as they learn about categories (e.g., Rehder, 2006, 2009, 2015; Rehder & Hastie, 2004; Rehder & Kim, 2010; Vogel, Kutzner, Freytag, & Fiedler, 2014). Given the central role that prediction plays in cognition, it is likely that subjective FOS may influence a wide array of judgments in such tasks.

6.2.3 Granularity of the effect vs. granularity of the predicted variable

In the present experiments, judgments of the objective strength of cause-effect relationships were influenced by the granularity of the effect, though not consistently by the granularity of the cause. One question is whether this finding is actually driven by this distinction between cause and effect. Another hypothesis is that it is actually driven by the granularity of the variable that the learner needs to predict from the cue that is presented. (In all the present experiments, the cause was the cue and the effect was the outcome being predicted.) This hypothesis could be tested with a design that manipulated whether participants inferred the state of the effect from the state of the cause (as in the present experiments), or inferred the state of the cause from the effect. This latter form of reasoning is referred to as “diagnostic”, in contrast to “predictive” reasoning, and prior research has documented asymmetries in how well people are able to reason in each direction (e.g., Fernbach, Darlow, & Sloman, 2011; Waldmann & Holyoak, 1992).

If subjective FOS is influenced only by the granularity of the predicted variable, then when participants learn about the causal relationship while making diagnostic predictions of the cause from the effect, there should be an influence of the cause’s granularity but not the effect’s granularity. On the other hand, reasoning from causes to effects (i.e. predictive reasoning) can be considered more natural because the variables are observed in a natural sequence; in diagnostic reasoning, learners infer the state of the cause, which happened prior to the observed effect. Thus, it is possible that the granularity of the effect has a privileged status in causal learning and may still influence learners’ judgments of causal strength.

6.3 The role of granularity in other tasks

The present experiments focused on the role of granularity in a learning task in which learners made predictions of a single effect based on observing a single cause, and finally estimated the objective strength of the causal relationship. There are other tasks that learners engage in when learning and reasoning about causal relationships. An open question of interest is whether the role of granularity is similar or different across various tasks.

For example, in some causal learning tasks, the goal is not to learn the causal strength underlying a known cause-effect relationship, but to learn the causal structure underlying a set of variables. To learn which variables exert a causal influence on others, learners make interventions on certain variables and observe what outcomes occur amongst the other variables. The majority of these experiments utilize binary variables (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Bramley et al., 2015; Lagnado & Sloman, 2004, 2006), although some have utilized continuous variables (Davis et al., 2018; Hagmayer et al., 2010; Soo & Rottman, 2014). However, none have investigated the role of the granularity of the variables. Learners in this task often try to identify causal variables that grant them maximal control over other variables (Coenen, Rehder, & Gureckis, 2015; also see Kim, Luhmann, Pierce, & Ryan, 2009). Given that strategy, learners may be more likely to infer that fine-grained variables are causes because such variables appear to grant learners more control over other variables, which would imply sensitivity to the specificity criterion (Franklin-Hall, 2016; Griffiths, et al., 2015; Weslake, 2010).

In some other causal learning tasks, learners experience data of multiple causes and a single effect, and the goal is to identify which cause or causes have an influence on the effect and the strength of each cause-effect relationship after controlling for the other causes (Coenen, Bramley, Ruggeri, & Gureckis, 2017; Derringer & Rottman, 2016, 2018; Spellman, 1996; Spellman, Price, & Logan, 2001). In past research, the granularities of all causes are held constant (typically, they are binary variables). Compared to the present experiments, such tasks require learners to attend to much more information due to there being multiple causes. In such situations, the feelings-of-success heuristic may not be applied because learners would have to make predictions from a combination of causes. Woodward (2010, 2018b) proposed the proportionality criterion as a way to evaluate if a particular variable is a good causal candidate for an effect, which is especially relevant for selecting the most appropriate causal candidate amongst others (also see Woodward, 2016). Thus, in a learning context involving multiple causes and a single effect, controlling for the causal strength of all causes, learners may be sensitive to the proportionality criterion and make stronger judgments for causes that are proportional to the effect.

It would also be interesting to investigate the role of the granularity of features used in categorization (Goldstone, 1994; Gregan-Paxton, Hoeffler, & Zhao, 2005; Kemp, 2012; Stewart & Chater, 2002). For example, research on causal categorization investigates how the causal relationships between the features of entities influence categorization (Ahn, 1999; Kim et al., 2009; Rehder, 2003; Rehder & Burnett, 2005; Rehder & Kim, 2006, 2010). In existing research of causal categorization, learners are typically presented with features represented with binary variables. However, existing research on function learning has investigated the learning of categories involving multi-level variables (e.g., DeLosh, Busemeyer, & McDaniel, 1997). While more fine-grained features provide more information from a statistical perspective and would theoretically allow learners to classify objects into more complex categories or permit better discrimination between categories, the present research suggests that coarse-grained features may lead to learners inferring stronger relationships between features that are coarse-grained, which may lead to stronger judgments of objects being members of a particular category.

6.4 Causal learning with a discretized variable when the underlying continuous variable is also observable

In the present experiments, I used a simple learning context in which participants viewed the states of discretized variables that plausibly represented a continuous cause and effect. Participants were told that the granularities of the discretized variables reflected the sensitivity of each variable's measurement (i.e. more sensitive measurements resulted in more fine-grained scales). In contrast to this simple learning context, people sometimes have access to both the state of a continuous variable and an accompanying discretized state (based on that continuous variable). For example, a doctor might measure a patient's blood glucose level and obtain a value on a continuous scale (e.g., 105 mg/dL). In addition, the measurement tool (e.g., medical software) might provide an accompanying label for that value (e.g., "low", "normal", or "high"). That label can be viewed as a level on a discretized scale.

Suppose that a learner is trying to estimate the causal strength of the relationship between a binary variable (e.g., patients who do vs. do not take a drug) on an outcome that is simultaneously represented as both a continuous and a discretized variable (e.g., mg/DL and labels of "low", "normal", or "high"). In such a situation, there are several interesting questions that arise. First, when learning the causal strength of the drug, an important question is whether learners would focus on the underlying continuous measure of blood glucose, or the accompanying discretized variable. This could be tested by presenting participants with datasets in which the correlation between the drug's dosage and the continuous measure of blood glucose diverges from the correlation between the drug's dosage and the accompanying discretized labels for the blood glucose level. Although the correlation between the drug's dosage and measurements of blood glucose on the continuous scale provides a more accurate estimate of the objective causal strength, learners may find attending to the discretized labels less cognitively demanding when trying to predict the blood glucose level (i.e. it is easier to predict if a patient's blood glucose is "too low" or "too high" compared to predicting an exact level on the continuous measure). Prior research has found that even when presented with continuous or ambiguous values, learners often assimilate those values into more coarse-grained categories for learning and classification (Marsh & Ahn, 2009; Murphy & Ross, 1999; Tajfel & Wilkes, 1963; Voss, Rothermund, & Brandtsta, 2008).

A second and related question of interest is what factors moderate whether learners attend to an underlying continuous measure of a variable or its accompanying discretized variable. Using the same example, learners may attend more to the continuous measure of blood glucose if they have reason to think the discretized cutoffs are arbitrary – e.g., if individual doctors make their own judgments of what level is "too high" as opposed to the cutoffs being agreed upon by medical community, if the underlying distribution is bimodal, or if the cutoff signifies a medically-relevant threshold (e.g., above a certain level, a secondary symptom occurs).

Finally, in contexts in which learners have access to an underlying continuous variable, a question of interest is how the granularity of an accompanying discretized variable influences judgments of causal strength. This could be tested by presenting participants datasets with identical observations of the underlying continuous variable, but varying the granularity of the accompanying discretized variable. If participants make implicit predictions of the discretized variable, they may feel the prediction task is easier when the discretized variable is coarse-grained, leading to stronger judgments of causal strength (due to the feelings-of-success heuristic). However, if the states of the underlying continuous variable are observable, perhaps the influence of the discretized variable's granularity would be attenuated.

6.5 Contributions

Initial research on causal learning focused on highly-simplified learning contexts involving binary variables (Allan & Jenkins, 1983; Cheng, 1997; Griffiths & Tenenbaum, 2005; Jenkins & Ward, 1965). Subsequent research has focused on more complex learning contexts by incorporating more fine-grained variables (e.g., Chow, Don, Colagiuri, & Livesey, 2018; Marsh & Ahn, 2009; Saito, 2015; Soo & Rottman, 2018; White, 2013).

The primary motivation of the present research was to investigate the role of the cause and effect's granularity in prediction and causal learning. The results show that at least the granularity of the effect consistently influenced learners' judgments of causal strength, opening up possibilities for how prior findings involving various tasks related to causal learning and reasoning may also depend on the granularity of the variables involved. Future research investigating the role of granularity in a wider range of cognitive tasks would provide a deeper understanding of how people reason and learn about causal relationships in complex environments.

Another theoretical contribution of the present research was providing an empirical test for concepts of causation from philosophy. While theories from philosophy are relevant to and inform our understanding of causation (e.g., Woodward, 2003), they are rarely subjected to empirical testing. The present research is an initial step towards a line of research that subjects philosophical theories of causation to testing via behavioral experiments, in line with the burgeoning field of experimental philosophy (Knobe, 2007; Rose & Danks, 2013; Sosa, 2007). While it is perhaps clear how philosophical theories can inform normative accounts of causal cognition (e.g., Cartwright, 2002, 2004; Danks, 2005; Eberhardt & Scheines, 2007; Glymour, 1998; Gopnik & Schulz, 2007; Lewis, 1973, 2000; Salmon, 1994; Spirtes, 2010; White, 1990; Woodward, 1996), it is less common for descriptive accounts based on experimental findings to inform philosophical accounts of causation. The present research aimed to integrate findings from psychology with the philosophy literature, hopefully leading to future research in a similar vein that fosters a broader interdisciplinary focus.

In addition to the theoretical contributions to causal learning research, there are practical implications of the present findings. People commonly have to make decisions based upon cues that may be more coarse- or fine-grained; for example, in medical diagnosis (Schwartz, Gorry, Kassirer, & Essig, 1973), and when making social judgments (e.g., Caruso, Mead, & Balcetis, 2009; Voss, Rothermund, & Brandtsta, 2008). The present research suggests that the granularity of the information people receive or encode can influence the judgments they make when learning about relationships between variables. Furthermore, the finding that the granularity of the effect influences estimates of objective causal strength suggests that discretizing variables (e.g., presenting people with simplified variables in the output of tests and decision support systems) can sometimes lead to biases in judgments.

6.6 Conclusion

Four experiments found that when learning about causal relationships while making predictions of an effect, learners judged causal strength to be greater when the effect was more coarse-grained, despite the objective causal strength being held constant. Through a mediation analysis, learners' subjective feelings-of-success when making predictions were identified as the mechanism by which the granularity of the effect influenced causal judgments.

These experiments paint an optimistic picture of human causal learning and reasoning; substituting subjective feelings-of-success for judgments of objective causal strength – which I refer to as the feelings-of-success heuristic – is sensible because in many contexts, stronger causal relationships can be used to make more accurate predictions. However, there are many real-world factors that can lead human learners into making errors; when subjective feelings-of-success are influenced by factors unrelated to the objective causal strength like the granularity of variables, using the feelings-of-success heuristic can lead to biases in judgment.

Appendix A Assessing the relationship between the cause and effect in the stimuli

In all datasets used as stimuli, regardless of the granularity of the cause or effect, the correlation between variables was $r = +.60 \pm .01$ (or $r = -.60 \pm .01$ in the reverse-coded versions). An additional property of the datasets was that the relationship between the variables in each dataset was only linear, rather than a higher-order polynomial.

When the cause had only two levels, only a linear relationship could be estimated between the cause and effect. However, when the cause had more levels – e.g., $\text{Levels}_C = 13$ – there could be a higher-order (e.g., quadratic) relationship between the cause and effect. In some datasets, both linear and quadratic functions might fit the data. To ensure that the function relating the cause and effect was consistent, I evaluated each generated dataset (see Section 2.1.3) using the BayesVarSel package in R (Garcia-Donato & Forte, 2018) to determine that the data was best fitted with a linear function.

For each dataset, I fitted a linear model that predicted the state of the effect (E) from the state of the cause (C), and a quadratic model that predicted E from C^2. Next, I used the BayesVarSel package to compute Bayes factor and the posterior probability of each model relative to a baseline model in which E is predicted only by an intercept, with a uniform prior probability across the baseline model and the model being tested.

I retained as stimuli only those datasets in which the posterior probability of the quadratic model was lower than the posterior probability of the linear model.

Appendix B Participants' prediction accuracy in the trial-by-trial prediction task

The purpose of the current section is to present participants' trial-level accuracy in the prediction task, indicating that participants were indeed learning about the causal relationships in the data during the prediction task. If participants were learning about the causal relationship, they would make increasingly accurate predictions as each scenario progressed.

I assessed the accuracy of participants' trial-by-trial predictions by computing the standardized squared error of all predictions made by participants for each trial. Participants' prediction accuracies were compared to predictions made by a model that made random predictions on each trial (the *random* model) and a model that made predictions using a series of regression models predicting each new state of the effect from all prior observed values within a particular dataset (the *ideal observer* model[9]). These models were fitted to the same datasets that participants actually viewed in the experiments.

The following figures display the average accuracy of participants' trial-by-trial predictions over the 20 trials of all scenarios within each condition of a particular experiment. The average standardized squared error of participants' predictions within each trial are plotted with black circles, which can be compared to the average standardized squared error of predictions by the random (white triangles) and ideal observer models (white squares). I have fitted log functions to the average predictions for each model so that the trends can be compared visually.

[9] For the first two trials, during which there were insufficient prior observations to make predictions using regression, the ideal observer model made random predictions.

B.1 Trial-by-trial accuracy in Experiment 1

The results for participants' average trial-level prediction accuracy across conditions in Experiment 1 are displayed in Figure 19.

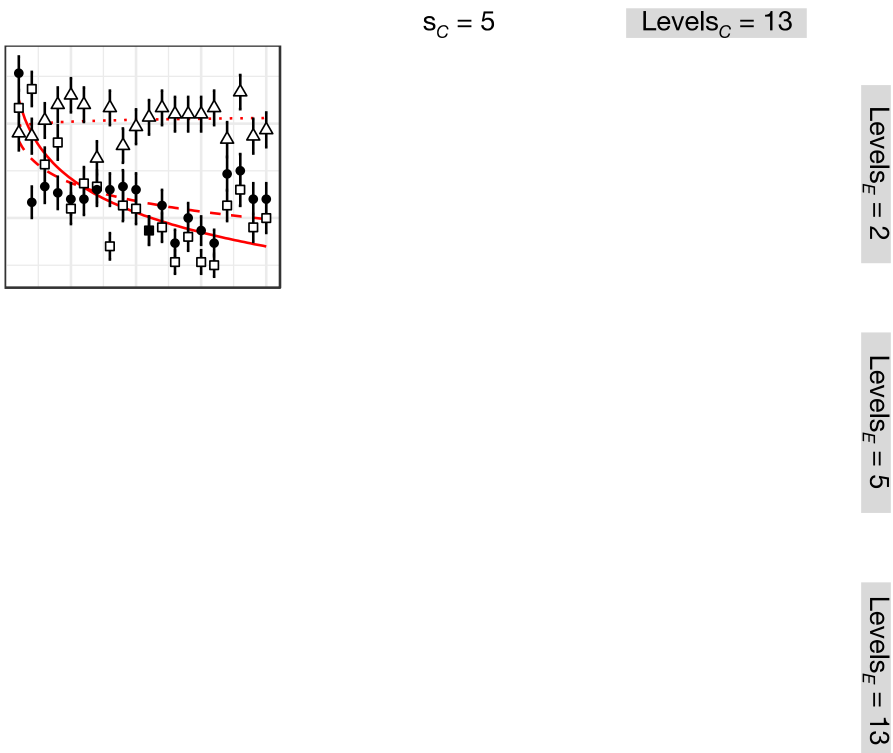

***Figure 19*. Average standardized squared error for predictions across trials by condition in Experiment 1. Error bars represent standard errors.**

On almost all trials, participants' predictions had lower standardized squared errors (i.e. were more accurate) than the random model but are slightly less accurate than predictions by the ideal observer model. Although participants' predictions always had higher standardized squared errors than the ideal observer model, it is evident that after sufficient trials, participants' learning of the function relating the cause and effect was better than chance, indicating participants were successfully learning about the causal relationship and successfully using the state of the cause to predict the state of the effect.

An issue with assessing prediction accuracy based on standardized squared error is that performance depends on the granularity of the effect (comparing across rows in Figure 19). When the effect has more levels, the standardized error is inflated because of the larger scale. This can be seen in the larger average standardized squared errors for conditions in which the effect has more levels (see the y-axis for plots across different rows). This was one motivation for introducing the accuracy bonus points scheme (Section 2.1.5) – to provide a cue for prediction accuracy that did not depend on the differing scales of the effect.

The crucial finding in the present section is that participants appeared to be learning the causal relationships from the data in all conditions, as demonstrated by their increasing accuracy (decreasing error) in predicting the state of the effect from observing the state of the cause.

B.2 Trial-by-trial prediction accuracy in Experiment 2

In Figure 20, I present results for Experiment 2, which include both the accuracy of predictions made by participants who received and did not receive bonus points. Ideally, participants should learn with similar accuracy both with and without feedback and incentives for making accurate predictions. The average standardized squared error of participants' predictions within each trial are plotted with black points (circles = participants who received bonus points, diamonds = participants who received no bonus points).

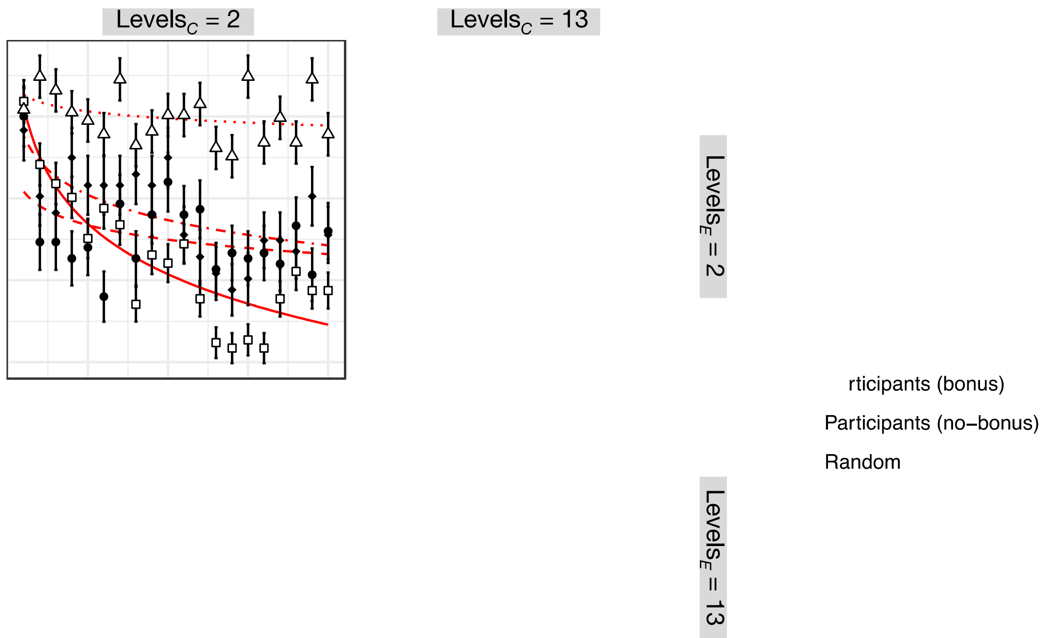

***Figure 20*. Average standardized squared error for predictions across trials by condition in Experiment 2. Error bars represent standard errors.**

On almost all trials, participants' predictions had lower standardized squared errors (i.e. are more accurate) than the random model but were slightly less accurate than predictions by the ideal observer model. Crucially, there did not appear to be differences in prediction accuracy between participants who received bonus points vs. those who received no bonus points. In both conditions, participants' predictions became more accurate with more trials, converging to levels somewhere between the accuracies of the random and ideal observer models.

B.3 Trial-by-trial prediction accuracy in Experiment 3

The results for participants' average trial-level prediction accuracy across conditions in Experiment 3 are displayed in Figure 21. In general, participants' trial-level prediction accuracy was similar to the prior experiments.

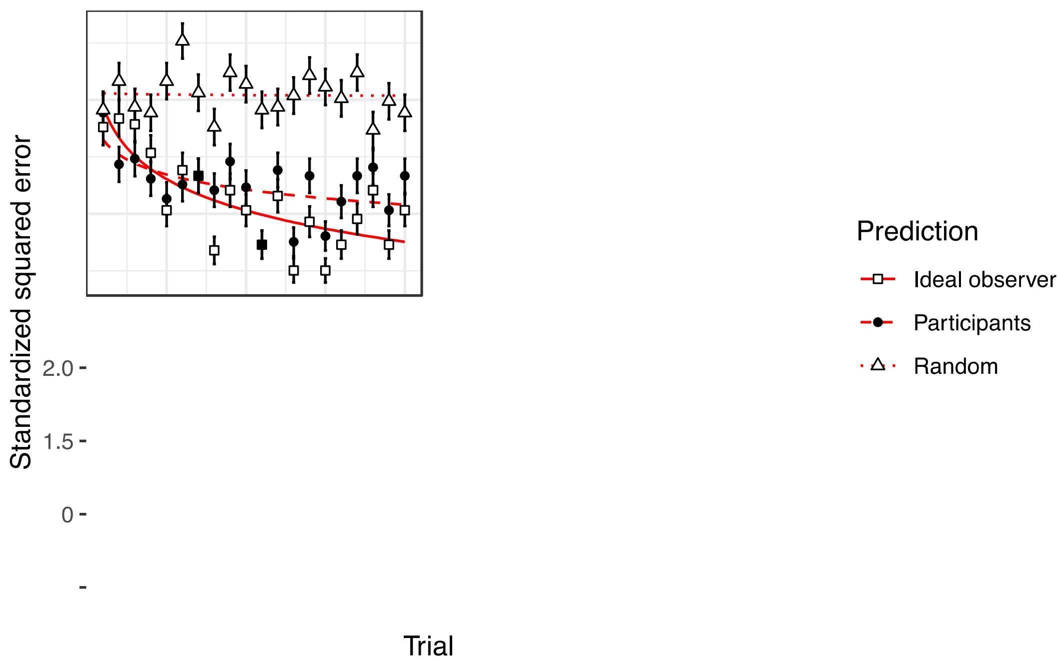

Figure 21. **Average standardized squared error for predictions across trials by condition in Experiment 3. Error bars represent standard errors.**

B.4 Trial-by-trial prediction accuracy in Experiment 4

The results for participants' average trial-level prediction accuracy across conditions in Experiment 4 are displayed in Figure 22. In general, participants' trial-level prediction accuracy was similar to the prior experiments. Note that in Experiment 4, only a third of participants in the prediction task condition actually made trial-by-trial predictions.

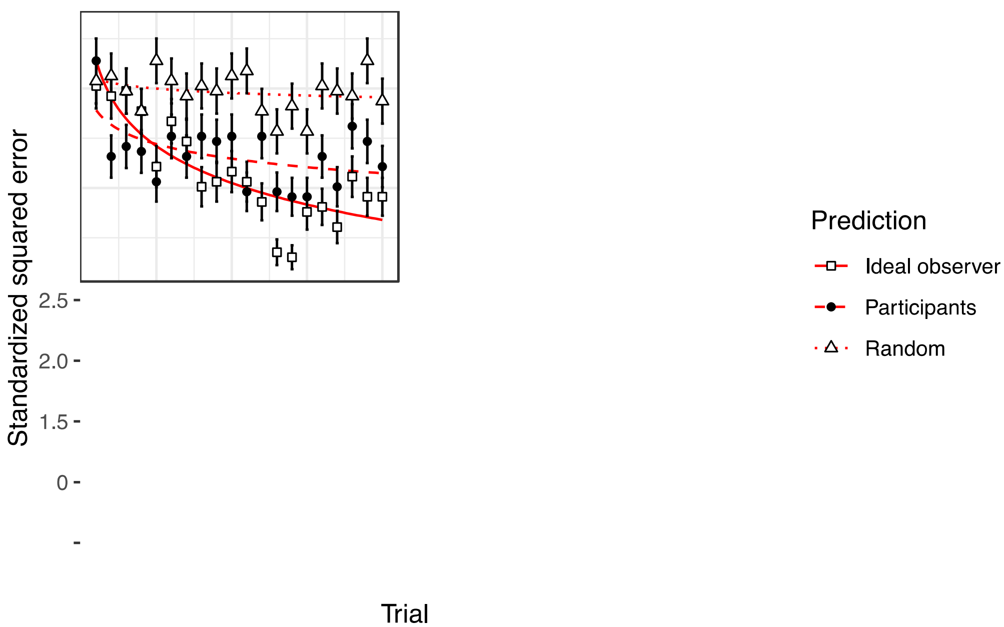

***Figure 22*. Average standardized squared error for predictions across trials by condition in Experiment 4. Error bars represent standard errors.**

Bibliography

Ahn, W. K. (1999). Effect of causal structure on category construction. *Memory & Cognition*, *27*(6), 1008–1023. http://doi.org/10.3758/BF03201231

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381–405. http://doi.org/10.1016/0023-9690(83)90024-3

Barberia, I., Baetu, I., Sansa, J., & Baker, A. G. (2014). When is a cause the "same"? Incoherent generalization across contexts. *The Quarterly Journal of Experimental Psychology*, *67*(2), 281–303. http://doi.org/10.1080/17470218.2013.804102

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. http://doi.org/10.1093/pan/mpr057

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology*, *53*(3), 155–167. http://doi.org/10.1016/j.jmp.2008.05.006

Birnbaum, M. H. (1976). Intuitive Numerical Prediction. *The American Journal of Psychology*, *89*(3), 417–429.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's Ship: Approximate Algorithms for Online Causal Learning. *Psychological Review*, *124*(3), 301–338.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in Causal Structure Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative Forgetful Scholars: How People Learn Causal Structure Through Sequences of Interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Breheny, P., & Burchett, W. (2017). Visualization of Regression Models Using visreg. *The R Journal*, *9*(2), 56–71.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*(6), 259–260.

Buehner, M. J. (2012). Understanding the past, predicting the future: causation, not intentional action, is the root of temporal binding. *Psychological Science*, *23*(12), 1490–7. http://doi.org/10.1177/0956797612444612

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119–40. http://doi.org/10.1037/0278-7393.29.6.1119

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Cartwright, N. (2004a). Causation: One Word, Many Things. *Philosophy of Science*, *71*(5), 805–820. http://doi.org/10.1086/426771

Cartwright, N. (2004b). From Causation To Explanation and Back. In B. Leiter (Ed.), *The Future for Philosophy* (pp. 230–246). Oxford: Oxford University Press.

Caruso, E. M., Mead, N. L., & Balcetis, E. (2009). Political partisanship influences perception of biracial candidates' skin tone. *Proceedings of the National Academy of Sciences*, *106*(48), 20168–20173.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. http://doi.org/10.1037//0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–82.

Chow, J. Y. L., Don, H. J., Colagiuri, B., & Livesey, E. J. (2018). Illusory causation and outcome density effects with a continuous and variable outcome. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1494–1499). Austin, TX: Cognitive Science Society.

Coenen, A., Bramley, N. R., Ruggeri, A., & Gureckis, T. M. (2017). Beliefs about sparsity affect causal experimentation. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1788–1793). Austin, TX: Cognitive Science Society.

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

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Crocker, J. (1981). Judgement of covariation by social perceivers. *Psychological Bulletin*, *90*(2), 272–292.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. http://doi.org/10.1016/S0022-2496(02)00016-0

Danks, D. (2005). The supposed competition between theories of human causal inference. *Philosophical Psychology*, *18*(2), 259–272. http://doi.org/10.1080/09515080500169371

Davis, Z. J., Bramley, N. R., & Rehder, B. (2018). Causal Structure Learning with Continuous Variables in Continuous Time. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 287–292). Austin, TX: Cognitive Science Society.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–86.

Derringer, C. J., & Rottman, B. M. (2016). Temporal causal strength learning with multiple causes. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 758–763). Austin, TX: Cognitive Science Society.

Derringer, C. J., & Rottman, B. M. (2018). How people learn about causal influence when there are many possible causes: A model based on informative transitions. *Cognitive Psychology*, *102*, 41–71. http://doi.org/10.1016/j.cogpsych.2018.01.002

Díez-Alegría, C., Vázquez, C., & Hernández-Lloreda, M. J. (2008). Covariation assessment for neutral and emotional verbal stimuli in paranoid delusions. *British Journal of Clinical Psychology*, *47*(4), 427–437. http://doi.org/10.1348/014466508X332819

Donkin, C., Rae, B., Heathcote, A., & Brown, S. D. (2015). Why is accurately labelling simple magnitudes so hard? A past, present and future look at simple perceptual judgment. In J. R. Busemeyer, Z. Wang, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 121–141). Oxford University Press.

Eberhardt, F., & Scheines, R. (2007). Interventions and Causal Inference. *Philosophy of Science*, *74*(5).

Erlick, D. E. (1966). Human estimates of statistical relatedness. *Psychonomic Science*, *5*(10), 365–366. http://doi.org/10.3758/BF03328441

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology*, *73*(1), 9–14. http://doi.org/10.1037/h0024138

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505. http://doi.org/10.1111/j.1469-8986.2007.00531.x

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

Finucane, M. L., Alhakami, A. L. I., Slovic, P., & Johnson, S. M. (2000). The Affect Heuristic in Judgments of Risks and Benefits. *Journal of Behavioral Decision Making*, *17*, 1–17.

Fisher, M., & Keil, F. C. (2018). The Binary Bias: A Systematic Distortion in the Integration of Information. *Psychological Science*, *29*(11), 1846–1858. http://doi.org/10.1177/0956797618792256

Fitzsimons, G. J. (2008). Death to Dichotomizing. *Journal of Consumer Research*, *35*(1), 5–8.

Forgas, J. P. (1995). Mood and judgment: The Affect Infusion Model (AIM). *Psychological Bulletin*, *117*(1), 39–66. http://doi.org/10.1037/0033-2909.117.1.39

Franklin-Hall, L. R. (2016). High-Level Explanation and the Interventionist's "Variables Problem." *British Journal for the Philosophy of Science*, *67*(2), 553–577. http://doi.org/10.1093/bjps/axu040

Garcia-Donato, G., & Forte, A. (2018). Bayesian Testing, Variable Selection and Model Averaging in Linear Models using R with BayesVarSel. *The R Journal*, *10*(1), 155–174.

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review*, *58*(6), 446–459. http://doi.org/10.1037/h0054482

Gelman, A., & Park, D. K. (2009). Splitting a Predictor at the Upper Quarter or Third and the Lower Quarter or Third. *The American Statistician*, *63*(1), 1–8. http://doi.org/10.1198/tast.2009.0001

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, *3*(1), 20–29.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100–121.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds And Machines*, *8*, 39–60.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, *109*(1), 75–90. http://doi.org/10.1037//0033-295X.109.1.75

Goldstone, R. L. (1994). Influences of Categorization in Perceptual Discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200.

Gopnik, A., & Schulz, L. (2007). *Causal Learning: Psychology, Philosophy and Computation*. Oxford University Press.

Gregan-Paxton, J., Hoeffler, S., & Zhao, M. (2005). When Categorization Is Ambiguous: Factors That Facilitate the Use of a Multiple Category Inference Strategy. *Journal of Consumer Psychology*, *15*(2), 127–140.

Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review*, *15*(2), 107–141. http://doi.org/10.1177/1088868310367640

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–71. http://doi.org/10.1037/a0020976

Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., Grif, P. E., … Knight, R. (2015). Measuring Causal Specificity. *Philosophy of Science*, *82*(4), 529–555.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–84. http://doi.org/10.1016/j.cogpsych.2005.05.004

Grzymala-Busse, A. (2011). Time Will Tell? Temporality and the Analysis of Causal Mechanisms and Processes. *Comparative Political Studies*, *44*(9), 1267–1297. http://doi.org/10.1177/0010414010390653

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 33–50. http://doi.org/10.1037/a0028643

Hagmayer, Y., Meder, B., Osman, M., Mangold, S., & Lagnado, D. (2010). Spontaneous Causal Learning While Controlling A Dynamic System. *The Open Psychology Journal*, *3*, 145–162. http://doi.org/10.2174/1874350101003010145

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, *138*(1), 22–38. http://doi.org/10.1037/a0014585

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*(7), 1128–37.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*, 139–183. http://doi.org/10.1016/S0166-4115(08)62386-9

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814. http://doi.org/10.1080/03640210701530755

Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*(1), 26–38. http://doi.org/10.1037/h0040013

Hertwig, R., Herzog, S. M., & Schooler, L. J. (2008). Fluency Heuristic: A Model of How the Mind Exploits a By-Product of Information Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1191–1206. http://doi.org/10.1037/a0013025

Hertzog, C., Dunlosky, J., Robinson, E. A., & Kidder, D. P. (2003). Encoding Fluency Is a Cue Used for Judgments About Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. http://doi.org/10.1037/0278-7393.29.1.22

Hogarth, R. M., & Karelaia, N. (2011). Heuristic and Linear Models of Judgment: Matching Rules and Environments. *Psychological Review*, *114*(3), 733–758. http://doi.org/10.1093/acprof:oso/9780199744282.003.0013

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–63. http://doi.org/10.1146/annurev.psych.121208.131634

Hutchinson, J. W., & Alba, J. W. (1997). Heuristics and biases in the “eyeballing” of data: The effects of context on intuitive correlation assessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 591–621. http://doi.org/10.1037//0278-7393.23.3.591

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology*, *25*(4), 652–665.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of Contingency Between Responses and Outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Johnson, S. G. B., & Keil, F. C. (2014). Causal Inference and the Hierarchical Structure of Experience. *Journal of Experimental Psychology: General*, *143*(6), 2223–2241.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 95–106.

Kareev, Y. (1995). Positive Bias in the Perception of Covariation. *Psychological Review*, *102*(3), 490–502.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–54. http://doi.org/10.1146/annurev.psych.57.102904.190100

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685–722. http://doi.org/10.1037/a0029347

Kim, N. S., Luhmann, C. C., Pierce, M. L., & Ryan, M. M. (2009). The conceptual centrality of causal cycles. *Memory & Cognition*, *37*(6), 744–58. http://doi.org/10.3758/MC.37.6.744

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, *16*(2), 85–125.

Knobe, J. (2007). Experimental Philosophy. *Philosophy Compass*, *2*(1), 81–92.

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, *12*(6), 467–472.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*(2), 416–428. http://doi.org/10.3758/MC.36.2.416

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*, 478–492. http://doi.org/10.1016/j.jml.2005.01.001

Koriat, A., & Nussinson, R. (2009). Attributing Study Effort to Data-Driven and Goal-Driven Effects: Implications for Metacognitive Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1338–1343. http://doi.org/10.1037/a0016374

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–76. http://doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–60. http://doi.org/10.1037/0278-7393.32.3.451

Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 640–649. http://doi.org/10.1037/0096-1523.11.5.640

Lejarraga, T., Dutt, V., & Gonzalez, C. (2010). Instance-based Learning: A General Model of Repeated Binary Choice. *The Journal of Behavioral Decision Making*, *25*(2), 143–153. http://doi.org/10.1002/bdm

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (2000). Causation as Influence. *Journal of Philosophy*, *97*(4), 182–197. http://doi.org/10.2307/2678389

Liljeholm, M. (2015). How multiple causes combine: independence constraints on causal inference. *Frontiers in Psychology*, *6*(1135), 1–12. http://doi.org/10.3389/fpsyg.2015.01135

Liljeholm, M., & Cheng, P. W. (2007). When Is a Cause the "Same"? Coherent Generalization Across Contexts. *Psychological Science*, *18*(11), 1014–1021.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–32. http://doi.org/10.1016/j.cogpsych.2010.05.002

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215. http://doi.org/10.3758/s13423-015-0808-5

Luhmann, C. C., & Ahn, W. K. (2007). BUCKLE: a model of unobserved cause learning. *Psychological Review*, *114*(3), 657–77. http://doi.org/10.1037/0033-295X.114.3.657

Luhmann, C. C., & Ahn, W. K. (2011). Expectations and interpretations during causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 568–87. http://doi.org/10.1037/a0022970

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods*, *7*(1), 19–40. http://doi.org/10.1037//1082-989X.7.1.19

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General*, *127*(3), 269–285. http://doi.org/10.1037/0096-3445.127.3.269

Marsh, J. K., & Ahn, W.-K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 334–52. http://doi.org/10.1037/a0014929

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate Median Splits and Spurious Statistical Significance. *Psychological Bulletin*, *113*(1), 181–190.

McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, *87*(1), 195–208. http://doi.org/10.1037/0033-2909.87.1.195

McClelland, G. H., Lynch, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false-positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*(4), 679–689. http://doi.org/10.1016/j.jcps.2015.05.006

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

McGrath, M. (1998). Proportionality and Mental Causation: A Fit? *Philosophical Perspectives*, *12*(Language, Mind, and Ontology), 167–176.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin*, *117*(3), 363–386.

Murphy, G. L., & Ross, B. H. (1999). Induction with cross-classified categories. *Memory & Cognition*, *27*(6), 1024–1041.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, *32*(5), 932–943. http://doi.org/10.1037//0022-3514.32.5.932

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference Fluency in Choice. *Journal of Marketing Research*, *44*(3), 347–356.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*(6), 1147–1152. http://doi.org/10.3758/BF03193104

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. http://doi.org/10.1016/j.tics.2008.02.014

Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. *Advances in Neural Information Processing Systems*, *24*.

Pachur, T., & Hertwig, R. (2006). On the Psychology of the Recognition Heuristic: Retrieval Primacy as a Key Determinant of Its Use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 983–1002. http://doi.org/10.1037/0278-7393.32.5.983

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. http://doi.org/10.1177/0963721414531598

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216. http://doi.org/10.1016/j.cogpsych.2013.09.002

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*(4), 577–596.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748. http://doi.org/10.1016/S0364-0213(03)00068-5

Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory & Cognition*, *34*(1), 3–16. http://doi.org/10.3758/BF03193382

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, *33*(3), 301–344. http://doi.org/10.1111/j.1551-6709.2009.01015.x

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. http://doi.org/10.1016/j.cogpsych.2014.02.002

Rehder, B. (2015). The Role of Functional Form in Causal-Based Categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 670–692. http://doi.org/10.1037/xlm0000048

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314. http://doi.org/10.1016/j.cogpsych.2004.09.002

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113–153. http://doi.org/10.1016/S0010-0277(03)00167-7

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 659–683. http://doi.org/10.1037/0278-7393.32.4.659

Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1171–1206. http://doi.org/10.1037/a0019765

Remko Duursma (2017). bootpredictlme4: Predict Method For lme4 With Bootstrap. R package version 0.1.

Rensink, R. A. (2017). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, *24*, 776–797. http://doi.org/10.3758/s13423-016-1174-7

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. E. Prokasy (Eds.), *Classical Conditioning II: Current Theory and Research* (pp. 64–99). New York: Appleton-Century-Crofts.

Restle, F., & Merr, C. T. (1968). An adaptation-level theory account of a relative-size iIIusion. *Psychonomic Science*, *12*(6), 229–230.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–276. http://doi.org/10.1016/j.actpsy.2007.05.004

Robb, D., & Heil, J. (2018). Mental Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition). Retrieved from https://plato.stanford.edu/entries/mental-causation/

Robinson, A. E., Sloman, S. A., Hagmayer, Y., & Hertzog, C. K. (2010). Causality in solving economic problems. *The Journal of Problem Solving*, *3*(1), 106–130. http://doi.org/10.7771/1932-6246.1081

Rose, D., & Danks, D. (2013). In defense of a broad conception of experimental philosophy. *Metaphilosophy*, *44*(4), 512–532.

Rottman, B. M. (2016). Searching for the Best Cause: Roles of Mechanism Beliefs, Autocorrelation, and Exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1233–1256.

Rottman, B. M., & Hastie, R. (2013). Reasoning About Causal Relationships: Inferences on Causal Networks. *Psychological Bulletin*. http://doi.org/10.1037/a0031903

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134. http://doi.org/10.1016/j.cogpsych.2016.05.002

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, *25*(4), 666–678. http://doi.org/10.1016/j.jcps.2015.04.004

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Saito, M. (2015). How People Estimate Effect Sizes: The Role of Means and Standard Deviations. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2075–2079). Austin, TX: Cognitive Science Society.

Salmon, W. (1994). Causality without Counterfactuals. *Philosophy of Science*, *61*(2), 297–312.

Sarbin, T. R. (1944). The logic of prediction in psychology. *Psychological Review*, *51*(4), 210–228. http://doi.org/10.1037/h0057400

Sarris, V. (1967). Adaptation-Level Theory: Two Critical Experiments on Helson's Weighted-Average Model. *The American Journal of Psychology*, *80*(3), 331–344.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79. http://doi.org/10.1016/j.cogpsych.2017.11.002

Schwartz, W. B., Gorry, G. A., Kassirer, J. P., & Essig, A. (1973). Decision Analysis and Clinical Judgment. *The American Journal of Medicine*, *55*, 459–472.

Sebanz, N., & Knoblich, G. (2009). Prediction in Joint Action: What, When, and Where. *Topics in Cognitive Science*, *1*, 353–367. http://doi.org/10.1111/j.1756-8765.2009.01024.x

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal Contiguity and the Judgement of Causality by Human Subjects. *The Quarterly Journal of Experimental Psychology*, *41B*(2), 139–159. http://doi.org/10.1080/14640748908401189

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. http://doi.org/10.1214/10-STS330

Sloman, S. A. (2005). *Causal Models: How People Think About the World and its Alternatives*. New York: Oxford University Press.

Sloman, S. A., & Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, *66*(3), 1–25. http://doi.org/10.1146/annurev-psych-010814-015135

Sloman, S. A., & Lagnado, D. A. (2005). Do We "Do"? *Cognitive Science*, *29*(1), 5–39. http://doi.org/10.1207/s15516709cog2901_2

Soo, K. W., & Rottman, B. M. (2014). Learning Causal Direction from Transitions with Continuous and Noisy Variables. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1485–1490). Austin, TX: Cognitive Science Society.

Soo, K. W., & Rottman, B. M. (2016). Causal learning with continuous variables over time. In A. Papafrogou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 153–158). Austin, TX: Cognitive Science Society.

Soo, K. W., & Rottman, B. M. (2018). Causal Strength Induction From Time Series Data. *Journal of Experimental Psychology: General*, *147*(4), 485–513.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Explorations*, *132*, 99–107. http://doi.org/10.1080/13869790701305905

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266–98. http://doi.org/10.1037/a0018620

Spellman, B. A. (1996). Acting as Intuitive Scientists: Contingency Judgments Are Made while Controlling for Alternative Potential Causes. *Psychological Science*, *7*(6), 337–342.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition*, *29*(2), 193–208. http://doi.org/10.3758/BF03194913

Spirtes, P. (2010). An introduction to causal inference. *Journal of Machine Learning Research*, *11*, 1643–1662. http://doi.org/10.2202/1557-4679.1203

Stewart, N. (2009). Decision by sampling: the role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*(6), 1041–1062. http://doi.org/10.1080/17470210902747112

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence Effects in Categorization of Simple Perceptual Stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3–11. http://doi.org/10.1037/0278-7393.28.1.3

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute Identification by Relative Judgment. *Psychological Review*, *112*(4), 881–911. http://doi.org/10.1037/0033-295X.112.4.881

Stewart, N., & Chater, N. (2002). The Effect of Category Variability in Perceptual Categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 893–907. http://doi.org/10.1037/0278-7393.28.5.893

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. http://doi.org/10.1016/j.cogpsych.2005.10.003

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. http://doi.org/10.1016/S0364-0213(03)00010-7

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (pp. 444–459). Cambridge, MA: MIT Press.

Suri, R., & Monroe, K. B. (2004). The Effects of Time Constraints on Consumers' Judgments of Prices and Products. *Journal of Consumer Research*, *30*, 92–104.

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, *3*, 9–44.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135–70.

Sutton, R. S., & Barto, A. G. (1987). A Temporal-Difference Model of Classical Conditioning. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society* (pp. 355–378). Austin, TX: Cognitive Science Society.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press. http://doi.org/10.1109/MED.2013.6608833

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology*, *52*(2), 101–114.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review*, *34*, 273–286.

Tingley, D., Yamamoto, T., Hirose, K., & Keele, L. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, *59*(5), 1–38.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Vallée-Tourangeau, F., Baker, A. G., & Mercier, P. (1994). Discounting in Causality and Covariation Judgments. *The Quarterly Journal of Experimental Psychology*, *47B*(2), 151–171. http://doi.org/10.1080/14640749408401354

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2016). Stable Causal Relationships are Better Causal Relationships. In A. Papafrogou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2663–2668). Austin, TX: Cognitive Science Society.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer.

Vogel, T., Kutzner, F., Freytag, P., & Fiedler, K. (2014). Inferring correlations: From exemplars to categories. *Psychonomic Bulletin & Review*, *21*, 1316–1322. http://doi.org/10.3758/s13423-014-0586-5

Voss, A., Rothermund, K., & Brandtsta, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology*, *44*, 1048–1056. http://doi.org/10.1016/j.jesp.2007.10.009

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian Conditioning: Application of a Theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and Learning* (pp. 301–336). London: Academic Press.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–27. http://doi.org/10.1037/0278-7393.31.2.216

Waldmann, M. R., & Hagmayer, Y. (2014). Causal Reasoning. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology*. New York: Oxford University Press.

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), 222–236. http://doi.org/10.1037/0096-3445.121.2.222

Well, A. D., Boyce, S. J., Morris, R. K., Shinjo, M., & Chumbley, J. I. (1988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory & Cognition*, *16*(3), 271–280.

Weslake, B. (2010). Explanatory Depth. *Philosophy of Science*, *77*, 273–294. http://doi.org/10.1086/651316

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, *108*(1), 3–18. http://doi.org/10.1037//0033-2909.108.1.3

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 710–727. http://doi.org/10.1037/0278-7393.29.4.710

White, P. A. (2013). Causal judgement from information about outcome magnitude. *The Quarterly Journal Of Experimental Psychology*, *66*(11), 2268–2288. http://doi.org/10.1080/17470218.2013.777750

White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *The Quarterly Journal of Experimental Psychology*, *68*(11), 2149–2174. http://doi.org/10.1080/17470218.2015.1009475

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society London B: Biological Sciences*, *358*(1431), 593–602. http://doi.org/10.1098/rstb.2002.1238

Woodward, J. (1996). Explanation, Invariance, and Intervention. *Philosophy of Science*, *64*, 26–41.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*, 287–318.

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. http://doi.org/10.1007/s11229-015-0810-5

Woodward, J. (2018a). Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese*. http://doi.org/10.1007/s11229-018-01998-6

Woodward, J. (2018b). Causal Cognition: Physical Connections, Proportionality, and the Role of Normative Theory. In W. Gonzalez (Ed.), *Philosophy of Psychology: Causality and Psychological Subject* (pp. 105-138). Berlin, Germany: De Gruyter.

Yablo, S. (1992). Mental Causation. *The Philosophical Review*, *101*(2), 245–280.

Zellner, A. (1988). Causality and causal laws in economics. *Journal of Econometrics*, *39*, 7–21.