COGNITIVE SCIENCE
A Multidisciplinary Journal

Stable Causal Relationships Are Better Causal Relationships

Nadya Vasilyeva,[a]* Thomas Blanchard,[b]* Tania Lombrozo[a]

[a]*Department of Psychology, University of California Berkeley*
[b]*Department of Philosophy, Illinois Wesleyan University*

Abstract

We report three experiments investigating whether people's judgments about causal relationships are sensitive to the robustness or *stability* of such relationships across a range of background circumstances. In Experiment 1, we demonstrate that people are more willing to endorse causal and explanatory claims based on stable (as opposed to unstable) relationships, even when the overall causal strength of the relationship is held constant. In Experiment 2, we show that this effect is not driven by a causal generalization's actual scope of application. In Experiment 3, we offer evidence that stable causal relationships may be seen as better guides to action. Collectively, these experiments document a previously underappreciated factor that shapes people's causal reasoning: the stability of the causal relationship.

Keywords: Stability; Robustness; Invariance; Sensitivity; Causality; Explanation; Background conditions; Moderating variables

1. Introduction

Consider two hypothetical cases. Case one: a medical journal reports an association between mutations in the BRCA1 gene and breast cancer. Case two: the same journal reports an equally strong association between mutations in the Gabrb1 gene and alcoholism. In both cases, the authors additionally report their results for participants of low

versus high socioeconomic status (SES). The relationship between the BRCA1 gene and breast cancer holds with the same strength for both subgroups. In contrast, the relationship between the Gabrb1 gene and alcoholism holds strongly in the low-SES subgroup, but nearly disappears in the high-SES subgroup. Would you be equally willing to endorse the following causal generalizations: that "mutations in the BRCA1 gene cause breast cancer" versus "mutations in the Gabrb1 gene cause alcoholism"?

While the overall association between putative cause and effect is equally strong in these two hypothetical cases, the causal associations differ in their *stability,* or insensitivity to background conditions (Woodward, 2006, 2010). Mutations in the Gabrb1 gene only elevate the risk of alcoholism under very specific conditions. In contrast, the relationship between mutations in the BRCA1 gene and breast cancer holds across multiple conditions. Do such differences in stability affect people's willingness to endorse causal and explanatory claims that invoke each of these relationships?

Stability is a well-known notion in the philosophical literature on causation. It was first introduced by Lewis (1986) under the name "insensitivity," and in contemporary work has been discussed most extensively by Woodward (2006, 2010). In Woodward's framework, one starts with a minimal notion of causal relevance, on which X is causally relevant to Y just in case X causally influences Y in at least *some* background circumstance b. Here X and Y are variables whose values represent event-types (in the case of a general causal claim) or the occurrence or non-occurrence of singular events (in the case of a singular causal claim). A background circumstance b is a situation that is not explicitly represented by X or Y, so that X's and Y's values do not specify whether b holds. On Woodward's view, causal influence is understood as counterfactual dependence under interventions, so that X is causally relevant to Y just in case in at least some background circumstance b, an intervention changing X's value would also change Y's value. With this framework in place, stability can then be defined as the extent to which the causal relationship $X \rightarrow Y$ holds in a variety of background circumstances. If $X \rightarrow Y$ holds in a wide variety of background circumstances—in particular, circumstances that we regard as "normal" or "important"—then it is relatively stable.[1]

Woodward argues convincingly that considerations regarding stability play an important role in scientific practice, especially in selecting appropriate levels of causal representation and explanation. For instance, stability considerations plausibly underlie certain features of explanatory practice in genetic psychiatry: Whether a genetic explanation of a mental disorder is appropriate depends in part on whether the causal relationship between the relevant gene and the disorder is stable—specifically, whether it is independent of the presence of other genes and of certain environmental and social factors (Woodward, 2010; see also Kendler, 2005).

Within psychology, researchers have also recognized notions of "invariance" or "robustness" as potentially relevant to how people identify causal relationships in the world (Sloman, 2005; Sloman & Lagnado, 2003). For example, Liljeholm and Cheng (2007) and Cheng, Liljeholm, and Sandhofer (2013) emphasize the importance of extracting invariant properties of causal relationships to justify causal generalizations. Evidence suggests that when the strength of a causal relationship is not invariant across contexts, people infer the presence of interacting background factors (Liljeholm & Cheng, 2007), and compute

separate estimates of "simple" and "conjunctive" (interactive) causal powers for a candidate and alternative cause (Cheng, 2000; Novick & Cheng, 2004). In our introductory example, people could use covariational evidence to infer that Gabrb1 mutations and SES influence alcoholism interactively, whereas BRCA1 mutations are a simple cause of breast cancer.

Prior work thus provides both theoretical reasons to expect stability to have an effect on causal generalizations, as well as empirical evidence that people possess important prerequisites to identifying stable relationships. However, this work has not investigated stability directly. The body of empirical research that is perhaps most relevant has investigated how people infer relationships of causal influence from patterns of covariation to develop a metric for "causal strength," such as ΔP (Allan, 1980) or power-PC (Cheng, 1997). Importantly, though, these metrics do not capture stability: ΔP, as originally proposed, captures the *average* strength of a causal relationship in a population (but see Spellman, 1996a,b, for an account of causal attribution based on computing ΔP over conditional contingencies), while causal power is typically calculated over subsets of data holding other variables constant (see Appendix S1 for more detail on how our proposal relates to handling of interactive causes and multiple informative focal sets within the power PC framework). Icard and Knobe (2016) offer an alternative metric that incorporates considerations of "robust [causal] sufficiency." But as they define it, the extent to which a cause $X = x$ is robustly sufficient for an effect $Y = y$ is measured by the probability of $Y = y$ given an intervention bringing about $X = x$, which in effect measures the average strength of the X-Y relationship in the population, not its stability. In contrast to these measures of causal strength, stability has to do with the extent to which a relationship holds *across* diverse segments of the population (or across various circumstances), specifically tracking how much the relationship varies from one segment to another. While causal strength and stability are often related (for instance, a deterministic causal relationship is a perfectly stable one), stability and causal strength can also come apart: In our example, the *BRCA1–breast cancer* relationship and the *Gabrb1–alcoholism* relationship are equally strong, but the former is more stable.

Another way to appreciate the difference between stability and strength is using the language of analyses of variance (ANOVAs). Prior work on causal learning has sought to characterize how people evaluate the existence or strength (effect size) of the conceptual equivalents of main effects, interactions, and simple effects in a factorial ANOVA. But stability is not reducible to any of these notions (see Appendix S1): The stability of a causal relationship, in the language of ANOVAs, tracks the extent to which the target cause interacts with background variables. Perhaps, a rough translation is that stability tracks how "qualified" a main effect is.[2]

Our project additionally departs from prior research in focusing not on the conditions that support an inference about the existence or strength of a causal relationship, but on the criteria that inform the evaluation of causal claims—and particularly causal generalizations. In many real-world cases, causal generalizations are made without specifying a variety of background circumstances: We say that lightning causes fire (without mentioning the necessary role of oxygen), that sex causes pregnancy (without specifying that the generalization is restricted to unprotected sex), or that aspirin reduces fever (without

conditioning this generic causal statement on patients' overall health condition, genotype, diet, age, etc.). Stability identifies one reason why such generalizations might be more or less appropriate, even once the underlying relationships of causal influence have been inferred. Returning to the example with which we began, our focus is not on the processes by which people might infer that SES is an element of an interactive cause of alcoholism, but on whether the instability of the Gabrb1 gene→alcoholism relationship with respect to SES affects the appropriateness of the generalization "mutations in the Gabrb1 gene cause alcoholism." Such causal generalizations are generic claims that abstract away from unspecified qualifications, much like "ducks lay eggs."

A handful of previous studies do provide evidence that causal claims are penalized for instability, but this evidence is indirect and potentially confounded with other factors. For instance, Lombrozo (2010) found that people are more willing to endorse a causal claim about a particular event (e.g., "Alice caused the music to start") when the candidate cause (a person) generated the effect (music) via a direct physical connection (throwing a ball at the "play" button) rather than via double prevention (preventing another person from unplugging a power cord). Moreover, when double prevention was involved, participants were more inclined to regard an agent as a cause of an outcome when the agent acted intentionally as opposed to accidentally. Lombrozo (2010) argues that both effects could be due to a difference in the stability of the relevant relationships: Both direct physical mechanisms and intentional actions will, in general, be more stable across variations in background circumstances. In the General Discussion, we consider other examples from prior research (Gerstenberg et al., 2012; Murray & Lombrozo, 2017; Nagel & Stephan, 2016; Phillips & Shaw, 2015). Crucially, though, no studies to date have investigated effects of stability while controlling other relevant features of the stimuli, such as the number of intermediate causes or the strength of the evidence supporting a causal relationship between a candidate cause and effect. To show that stability has an effect over and above causal strength, it is essential to consider cases for which stability varies while causal strength (e.g., measured as ΔP or causal power) is held fixed.

We conducted three experiments to investigate whether people are sensitive to stability. Participants were presented with evidence suggesting either that a causal relationship holds in only one out of two circumstances, or that it holds in both circumstances. The causal strength of the relationship (for the full set of cases) was held fixed across the two stability conditions. If people's causal and explanatory judgments are sensitive to stability considerations, this should be reflected in a lower willingness to say that C causes or explains E when the relationship holds only in one possible circumstance.

2. Experiment 1

The main goal of Experiment 1 was to examine the effect of stability on judgments of causal relationships when the causal strength of the relationships is held constant. To do so, we presented participants with evidence suggesting that a factor C has a causal influence on an effect E in a certain population. We further specified that some

members of the population had a certain property B (e.g., a behavioral or environmental characteristic) that other members of the population lacked. Participants were assigned to one of two conditions. In the *non-moderated* condition, participants were presented with further evidence suggesting that C has a causal influence on E both when B is present and when it is absent. In the *moderated* condition, by contrast, the evidence suggested that C causes E *only* when B is present (i.e., in the presence of the *enabling circumstance*). The causal strength (ΔP or power PC) of C→E in the overall population was the same in both conditions, but the stability of the relationship varied. The relationship was stable with respect to the *moderator variable* (presence or absence of B) in the non-moderated condition, but unstable with respect to this variable in the moderated condition. As the causal strength of the relationship was the same in both conditions, greater endorsement of a causal generalization in the stable (*non-moderated*) than in the unstable (*moderated*) condition would reveal an effect of stability.

In philosophy, the notion of stability has been applied to causal relations both between *types* (Woodward, 2010) and between *token* events (Woodward, 2006), and it is held to be important both for *causal* and *explanatory* judgments (Woodward, 2010). To examine the generality of the effects of stability, participants were thus asked to rate statements about the relationship between C and E in either the overall population or for specific token cases, and in the form of either causal or explanatory statements. We also anticipated that stability might be more important in evaluating explanatory claims, given both theoretical and empirical work suggesting a close connection between explanation and generalization, which stability is taken to support (e.g., Lombrozo & Carey, 2006; Vasilyeva & Lombrozo, unpublished data).

2.1. Method

2.1.1. Participants

A total of 182 participants were recruited on Amazon Mechanical Turk in exchange for $1.50. In all experiments, participation was restricted to users with an IP address within the United States and an approval rating of at least 95% based on at least 50 previous tasks. An additional 49 participants were excluded for failing a comprehension check (explained below).

2.1.2. Materials, design, and procedure

Participants first completed a short training to ensure that they could interpret covariation tables and were then placed in the role of a scientist (zoologist, botanist, geologist, or ornithologist) studying several natural kinds on a fictional planet. Table 1 shows the four kinds—zelmos, drols, grimonds, and yuyus—each associated with a triad of variables (putative cause, effect, and moderator). We illustrate the procedure with zelmos, but the structure was matched across cases.

The scientist was described as investigating the hypothesis that eating yona plants is causally related to developing sore antennas (see Appendix S2 for a sample vignette). Participants were told that to test the hypothesis, the scientist performed an experiment,

Table 1
Materials used in Experiments 1 (all four items) and 2 (zelmo and drol items only)

Item	Zelmo (lizard-like species)	Drol (mushroom)	Grimond (mineral)	Yuyu (bird)
Cause variable	Eating yona plants	Saline soil	Exposure to sulfuric acid	Eating marine snails
Effect variable	Sore antennas	Bumpy stems	Surface cracks	Brownish feather tint
Moderator variable	Drinking water (salty vs. fresh)	Exposure to forest fire smoke (occurred vs. not occurred)	Temperature (hot vs. cold)	Inhaling volcanic ash (occurred vs. not occurred)

selecting a random sample of 200 zelmos and randomly assigning them to two equal groups that ate a diet either containing or not containing yonas. Participants saw the results of the experiment in the form of a 2×2 covariation table cross-classifying zelmos based on whether they ate yonas or not, and whether they developed sore antennas or not (see Fig. 1a). The numbers in the table were selected to provide support for a relationship with causal strength equal to a ΔP of about .4 (range 0.39–0.42).

The scientist then decided to conduct a second experiment with a new, larger sample of 400 zelmos, again randomly assigning zelmos to one of the two diets. But this time the scientist discovered after the experiment that due to a miscommunication between research assistants, half of the zelmos were given salty water, and the other half were given fresh water. The two values of this potentially moderating variable were always said to occur normally on the planet; for example, in the wild, zelmos drink either fresh or salty water, depending on what's available. (This moderating variable played the role of a "background circumstance" relative to which the cause-effect relationship (e.g., eating yonas→sore antennas) was stable or unstable.) Luckily for the scientist, the moderator and cause variables varied orthogonally. Participants were told that "to see whether drinking salty water made a difference to the effects of yonas on sore antennas, you decide to look at the results of the experiment within each of these two groups." This time participants were presented with the data split into two tables, one for the salty water subgroup, and one for the fresh water subgroup, each table cross-classifying zelmos in terms of diet and antenna soreness (see Fig. 1).

Depending on condition, the split tables indicated a relationship that was either *moderated* or *not moderated*. In the moderated cases (illustrated in Fig. 1c), in one subgroup (salty water) the relationship between eating yonas and sore antennas was very strong ($\Delta P = .81$–0.86), while in the other subgroup (fresh water), the relationship disappeared ($\Delta P = .00$–0.01). In the non-moderated cases (Fig. 1b), each of the split tables corresponded to relationships with a ΔP comparable to the ~0.40 from the original, unsplit table. Importantly, the average strength of the relationship across the two split tables was
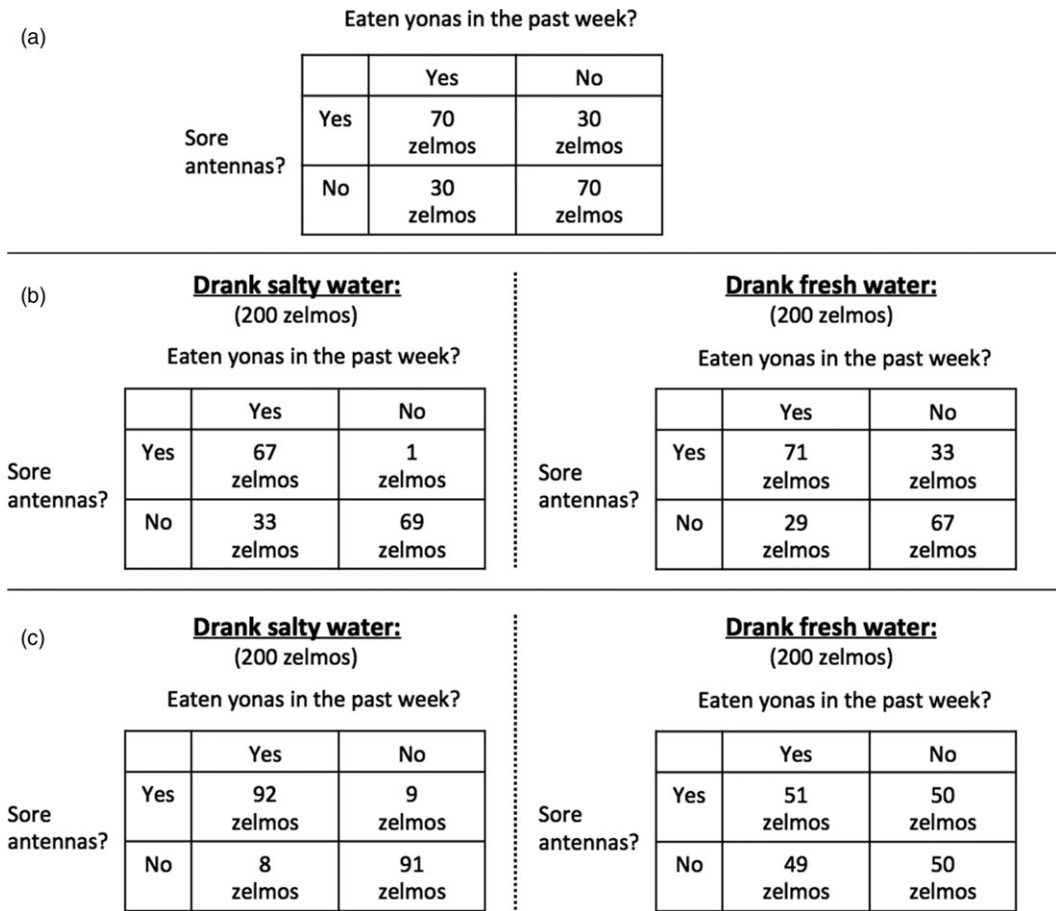
Fig. 1. Sample covariation matrices from Experiment 1: (a) original unsplit table, common across the moderated and non-moderated conditions; (b) split tables in the non-moderated condition, ΔP's = .36 and .38 (M = .37); (c) split tables in the moderated condition, ΔP's = .83 and 0.01 (M = 0.42).

the same in the moderated and non-moderated conditions (or differed by no more than 0.05 ΔP units, always in the direction working *against* our hypothesis[3]), and equaled the strength of relationship in the first table that participants saw for each item (within .03 ΔP units). The split tables were accompanied by a note for moderated [non-moderated] conditions: "The tables reveal that the data pattern looks very *different* [*similar*] for zelmos who drank salty water during the experiment and for zelmos who drank fresh water during the experiment. Please compare the two tables to see how different [similar] the patterns are."

Once all three covariation tables had been presented, participants evaluated either claims about *causal relationships* or *explanations* (Table 2). Each claim was presented either at the *type* or *token level*. All claims were unqualified; that is, they stated a relationship between eating yonas and sore antennas without mentioning the kind of water

the zelmo(s) in question drank.[4] In addition, participants evaluated one counterfactual statement for each scenario; for example, after learning about a group of zelmos who were fed yonas, drank salty water, and developed sore antennas, participants rated their agreement with the statement that "had these zelmos eaten yonas *but not drunk salty water*, their antennas would still have become sore." This statement was included to verify that participants differed across the moderated and non-moderated conditions in the role they attributed to the moderator.

At the end of the experiment, participants answered two multiple-choice comprehension check questions about each scenario they had read (e.g., "According to what you read, as a scientist on planet Zorg you were interested in evaluating the following hypothesis about zelmos: a. eating yona plants produces antenna soreness; b. eating drol mushrooms produces antenna soreness; c. eating mushrooms with stem bumps produces spotted antennas; d. antenna soreness makes zelmos eat yonas"). Participants who answered either question incorrectly were excluded from further analyses.

Across items, each participant saw two moderated cases and two non-moderated cases, presented in random order. Thus, Experiment 1 had a 2 moderator (moderated vs. non-moderated relationship) × 2 judgment (causal vs. explanatory) × 2 target (type vs. token) mixed design, with moderator manipulated within-subjects. The dependent variables were agreement with causal or explanatory claims, and agreement with counterfactual claims, measured on a 1 (strongly disagree) to 7 (strongly agree) scale.

2.2. Results and discussion

2.2.1. Causal and explanation ratings

Our main question was whether relationships with known moderators support causal and explanatory claims to the same extent as relationships without known moderators. A

Table 2
Sample causal and explanation judgments in Experiment 1, as a function of judgment type (causal vs. explanatory) and target (token vs. type)

	Causal Judgment	Explanation Judgment
Token	Your assistants select one of the zelmos with sore antennas from your second experiment. They call him Timmy. During the experiment Timmy has eaten yonas. You do not know whether Timmy drank fresh water or salty water during the experiment.	
	How much do you agree with the following statement about what caused Timmy's sore antennas? *Eating yonas caused Timmy's antennas to become sore*	How much do you agree with the following explanation of why Timmy has sore antennas? *Timmy's antennas became sore because he ate yonas*
Type	How much do you agree with the following statement about what causes zelmos' antennas to become sore? *For zelmos, eating yonas causes their antennas to become sore*	How much do you agree with the following explanation of why zelmos' antennas become sore? *For zelmos, antennas become sore because of eating yonas*

2 moderator (moderated relationship, non-moderated relationship) × 2 judgment (causal, explanatory) × 2 target (type, token) mixed ANOVA on causal and explanatory ratings revealed a main effect of moderator: as shown in Fig. 2a, participants were significantly less likely to agree with claims about causal and explanatory relationships when a relationship was moderated than non-moderated, $F(1, 178) = 163.22$, $p < .001$, $\eta_p^2 = 0.478$, even though moderated and non-moderated relationships were equated for overall strength (defined as the degree of covariation between putative causes and effects). There were no other significant main effects nor interactions,[5] suggesting that the effect of moderator was not itself moderated by the nature of the judgment (causal or explanatory, type or token). We thus failed to find support for the idea that stability might be a more important consideration in making explicitly explanatory claims relative to causal claims.

2.2.2. Counterfactual ratings

The counterfactual ratings showed a similar pattern: a 2 moderator (moderated relationship, non-moderated relationship) × 2 target (type, token) mixed ANOVA on counterfactual ratings showed a main effect of moderator, $M_{\text{non-mod}} = 5.48$, $SD = 1.19$, $M_{\text{mod}} = 3.41$, $SD = 1.21$, $F(1, 180) = 248.66$, $p < .001$, $\eta_p^2 = 0.580$, but no effect of target, $F(1, 180) = .25$, $p = .619$, and no interaction, $F(1, 180) = .22$, $p = .638$. This confirms that participants understood the enabling role of the moderator variable in the moderated condition.

Thus, the results of Experiment 1 indicate an effect of stability over and above causal strength: Holding causal strength fixed, causal claims are penalized for instability. The effect was consistent across tasks, holding for both causal and explanatory judgments at both the type and token levels.
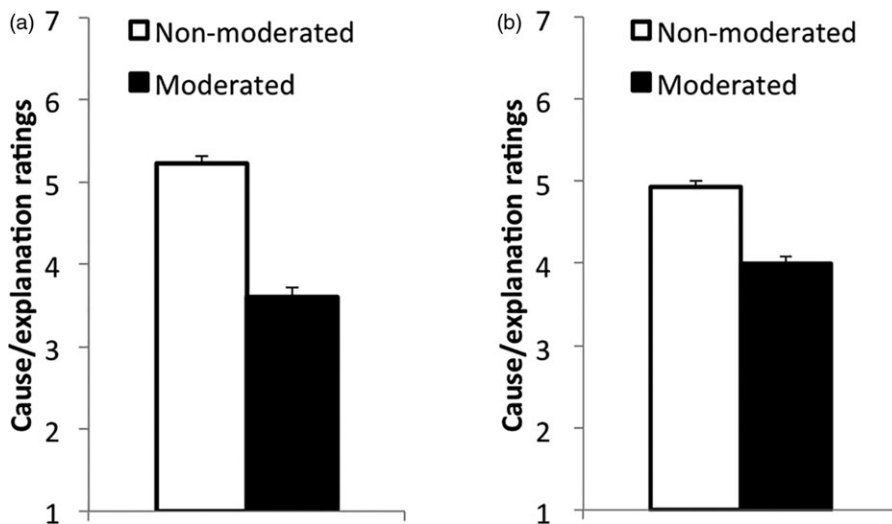

Fig. 2. The effect of moderator on ratings of causal and explanatory relationships in Experiment 1 (a) and Experiment 2 (b). Error bars correspond to 1 SEM.

3. Experiment 2

The results of Experiment 1 provide initial evidence for an effect of stability on causal and explanatory ratings. Yet these results are also amenable to alternative interpretations. In Experiment 1, the moderated relationship had two characteristics. First, it was relatively unstable, in that it held in only one circumstance described in the fictional world. The non-moderated relationship, by contrast, held in both circumstances. Second, the moderated relationship had a narrower *actual scope*, that is, the actual proportion of the population for which it held was relatively small: As the moderating variable took the value favoring the presence of the causal relationship in half of the actual members of the population, the moderated relationship held for only 50% of the actual population, and was clearly absent in the remaining 50% of the population. By contrast, the non-moderated relationship held in the entire actual population.

Because stability and actual scope covaried in Experiment 1, it could be that participants penalized unstable relationships, as we have suggested, or instead that they penalized relationships with narrow actual scope. Admittedly, this alternative explanation makes more sense for type causal statements than for token causal statements, which explicitly single out a specific individual. Yet it may be that evaluations of token causal statements of the form "*c* caused *e*" are sensitive to the actual scope of the causal generalization under which the *c-e* relationship falls. Either way, we thought it prudent to test this alternative, as stability and actual scope are not only conceptually distinct, but also dissociable. For instance, an unstable relationship can have wide actual scope if the circumstance in which it holds happens to be frequent (e.g., lots of zelmos happen to drink salty water).

To address the possibility that our results were driven by sensitivity to actual scope rather than stability, Experiment 2 not only varied the number of circumstances in which a causal relationship holds (thus investigating effects of stability) but orthogonally varied the relative size of the two subsets of the population broken down by the moderator variable (thus, varying actual scope). In Experiment 1, the proportion of the population for which the enabling circumstance (e.g., drinking salty water) held was always 50%, and therefore fixed actual scope to 50% of the population. In Experiment 2, we introduced two additional conditions: a *high-frequency* condition in which the enabling circumstance was present in 70% of the population and a *low-frequency* condition in which the enabling circumstance was present in 30% of the population. The actual scope of the moderated relationship thus varied across frequency conditions, but its (in)stability remained the same: In all frequency conditions, there was one possible circumstance (e.g., drinking fresh water) in which the causal relationship did not hold.

Experiment 2 also included a set of ratings concerning the structure and strength of causal relationships. Following past research on causal inference (Griffiths & Tenenbaum, 2005), participants were asked whether in their view a causal relationship between the cause (e.g., eating yonas) and effect (e.g., sore antennas) is *likely to exist*, and if so how

strong it is. We included these questions to test for effects of stability on these more familiar causal judgments, and also to determine whether effects of stability might be restricted to the kinds of causal and explanatory generalizations used in our agreement measures, which are presumably more susceptible to pragmatic effects concerning the omission of information about the value of the moderating variable (e.g., the failure to mention salty water in a causal or explanatory claim).

3.1. Method

Three-hundred-and-ninety-three participants (excluding an additional 83 participants who failed a memory check) were recruited on Amazon Mechanical Turk in exchange for $1.30.

3.1.1. Materials, design, and procedure

The materials, design, and procedure were the same as in Experiment 1, with the following exceptions. First, in the cover story, we presented split data tables in the context of an additional experiment designed to determine whether the moderator makes a difference (rather than a consequence of mistakes made by research assistants), and we increased the sample sizes in the hypothetical experiment to accommodate the changes in our design.

Second, we varied the *moderator frequency*: the base rate of the enabling circumstance (i.e., the moderator value under which the causal relationship held) in the natural population and in the sample. This circumstance (e.g., drinking salty water) occurred in either 30% (*low frequency*), 50% (*medium frequency*), or 70% of cases (*high frequency*). All numerical information was communicated to participants with descriptive text accompanied by intuitive illustrations representing the hypothetical study design (see Fig. 3 for a sample illustration from a high-frequency condition, and see Appendix S3 for a sample vignette). Participants were told that the sizes of the groups were intentionally matched to the frequency of the enabling circumstance in the natural population. Critically, we kept the mean strength of causal relationships averaged across each pair of split tables the same ($\Delta P = 0.31$) in the moderated and non-moderated condition across all frequency conditions.[6]

Third, participants answered additional questions about the structure and strength of causal relationships between pairs of variables (see Griffiths & Tenenbaum, 2005). For instance, a structure judgment might ask: "In your opinion, how likely is it that there is some causal relationship between eating yonas and having sore antennas?," rated on a scale from not at all likely (1) to very likely (7). A strength judgment might ask: "If there is a causal relationship between eating yonas and having sore antennas, how strong do you think it is?", rated on a scale from very weak relationship (1) to very strong relationships (7). Only participants who gave a rating higher than 1 in response to *structure* were asked to rate *strength*. Participants made such structure and strength judgments about the candidate cause and effect (e.g., eating yonas→sore antennas). For completeness and as a manipulation check, participants also made judgments about the moderator variable and

effect (e.g., drinking salty water→sore antennas). To prevent participant fatigue given these additional ratings, the number of items was reduced to two (see Table 1).

Thus, Experiment 2 had a 2 moderator (moderated vs. non-moderated relationship) × 2 judgment (causal vs. explanatory) × 2 target (type vs. token) × 3 moderator frequency (low 30%, medium 50%, high 70%) mixed design, with moderator manipulated within-subjects. We measured agreement with causal or explanatory claims and endorsement of causal structure and strength claims.

3.2. Results

Given that in Experiment 1 we did not find effects of target (type, token) or judgment (causal, explanatory), and they did not interact with moderator (the variable of main theoretical interest), the analyses we report in the main text for subsequent experiments are collapsed across target and judgment. The cases in which including these variables made a difference are reported in endnotes.

3.2.1. Causal and explanation ratings

A 2 moderator (moderated relationship, non-moderated relationship) × 3 moderator frequency (low, medium, high) mixed ANOVA on main causal and explanation ratings revealed that, as predicted, moderated relationships were rated lower than non-moderated
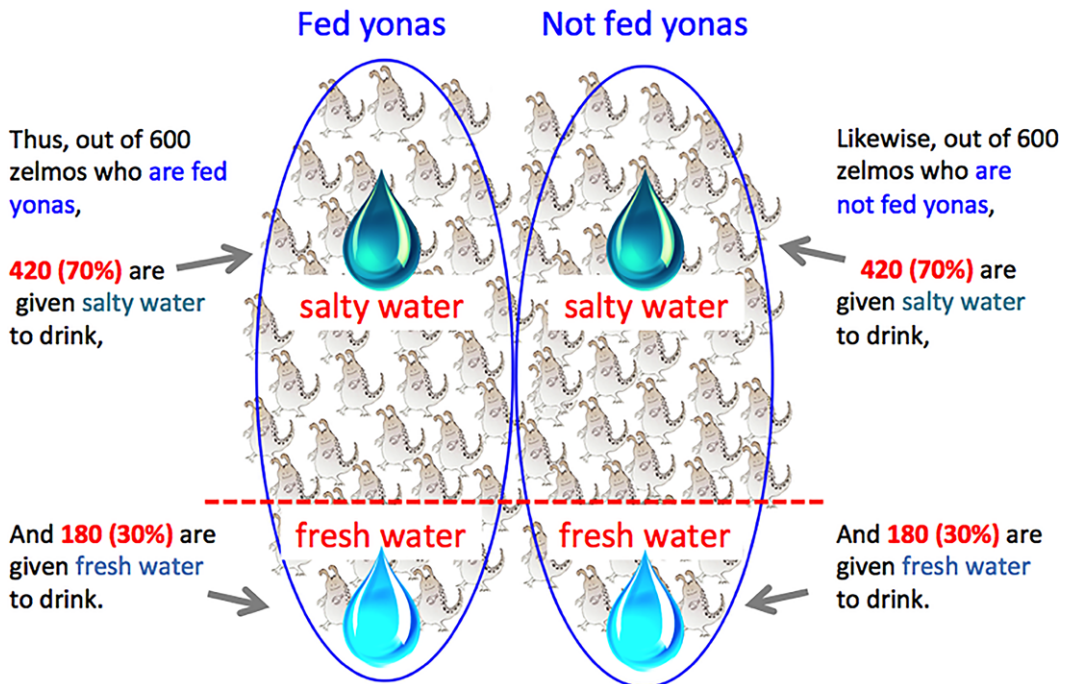

Fig. 3. Sample diagram provided to participants to illustrate the design of a hypothetical study (high-frequency moderator condition).

relationships, $F(1, 390) = 81.15$, $p < .001$, $\eta_p^2 = 0.172$, replicating the moderator effect from Experiment 1 (see Fig. 2b). Importantly, there was no interaction between moderator and frequency, $F(2, 390) = 1.85$, $p = .159$ (see Fig. 4a), suggesting that the effect attributed to stability in Experiment 1 is unlikely to reflect actual scope. There was also no main effect of frequency, $F(2, 390) = 1.59$, $p = .205$.[7]

3.2.2. Ratings of causal structure and strength

A 2 moderator (moderated relationship, non-moderated relationship) × 3 moderator frequency (low, medium, high) mixed ANOVA on causal structure ratings revealed a significant main effect of moderator, $F(1, 390) = 57.46$, $p < .001$, $\eta_p^2 = 0.128$: Ratings were on average higher for the non-moderated relationship than for the moderated relationship. As shown in Fig. 4b, there was no interaction between moderator and frequency, $F(2, 390) = 1.70$, $p = .183$, and no main effect of frequency, $F(2, 390) = 2.13$, $p = .121$.[8]

To analyze causal strength ratings across the full sample, trials for which a participant was not asked to rate strength because the structure question received a rating of 1 (not at all likely) were assigned a causal strength rating of zero. This effectively transformed strength ratings into an eight-point scale, 0–7. A 2 moderator (moderated relationship, non-moderated relationship) × 3 moderator frequency (low, medium, high) mixed ANOVA on transformed causal strength ratings revealed a significant main effect of moderator, $F(1, 389) = 39.43$, $p < .001$, $\eta_p^2 = 0.092$, with higher ratings for non-moderated than moderated relationships. There was no main effect of frequency, $F(2, 389) = 0.58$, $p = .560$, but the moderator by frequency interaction approached significance, $F(2, 389) = 2.49$, $p = .084$. As Fig. 4c shows, this marginal interaction was driven by variation in ratings for *non*-moderated relationships (simple effects: low vs. medium $p = .046$, medium vs. high $p = .053$, low vs. high $p = .951$), rather than variation in ratings for the moderated relationships (all simple effect ps $\geq .465$). Tests of simple effects revealed that the effect of moderator held reliably in all three frequency conditions ($p_{\text{low}} = .004$, $p_{\text{medium}} < .001$, $p_{\text{high}} = .012$).[9]

As expected, the additional structure and strength ratings confirmed that participants recognized the causal relevance of the moderating variable in the moderated condition.[10]

3.3. Discussion

Across a variety of judgments (type- and token-level causal and explanatory claims, causal structure ratings, and causal strength ratings), we observed a robust effect of stability: non-moderated relationships were rated higher than moderated relationships, even though these relationships were matched for average strength. Importantly, this pattern of results could not be explained by variation in the actual scope of the relationship being evaluated: Actual scope did not have a reliable effect (at least for the range of values tested), whereas the effect of stability emerged for each frequency condition.

Before moving on, we consider four additional explanations for the effects we attribute to stability. First, it could be that participants reduced their ratings in the moderated condition not because they received evidence of moderation *per se*, but because they
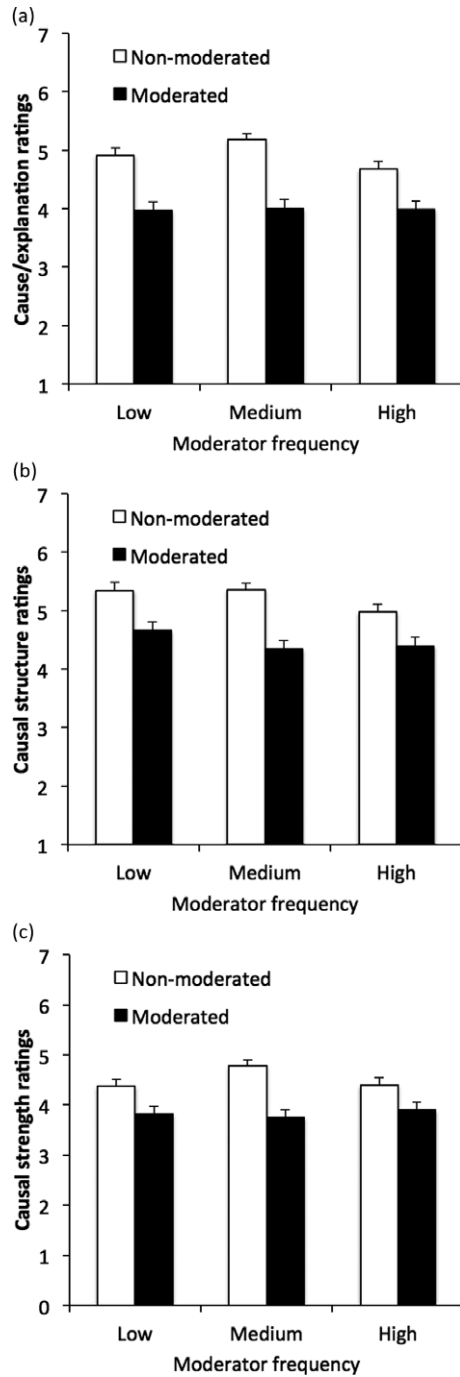
Fig. 4. Ratings of causal and explanatory claims (a), causal structure (b) and causal strength (c) of candidate cause–effect relationships as a function of relationship stability (non-moderated, moderated) and moderator frequency (low, medium, high). Error bars correspond to 1 SEM.

received evidence that their initial causal conclusions (on the basis of the initial covariation table) were inaccurate or incomplete, thus reducing their confidence relative to participants in non-moderated condition. To test this possibility, we ran an additional experiment, reported in Appendix S4, in which participants received evidence of an additional, independently sufficient cause (instead of evidence for the causal instability of the target relationship). Receiving this information did not lower ratings for the target causal generalization relative to a condition in which evidence for the alternative cause was not presented, suggesting that simply learning any new information about a causal system is insufficient to produce the effect we attribute to stability.

Second, it could be that our effects of moderation do not reflect stability as such, but instead the way in which causal power is computed over subpopulations. Fortunately, this possibility can be evaluated on the basis of data gathered in Experiment 2. In Experiments 1 and 2, moderated and non-moderated relationships were equated for average strength (ΔP, causal power) calculated over summary covariation tables (i.e., collapsing across the two subpopulations defined by the moderator variable, e.g., salty water and fresh water). If participants evaluated causal strength on the basis of this summary table (equivalent to "contextual power" as defined by Cheng, 2000), then it is clear that the effects of stability observed in Experiments 1 and 2 cannot be reduced to causal strength, which did not vary across the moderated and non-moderated conditions. However, an alternative possibility is that participants instead calculated a measure of causal strength for the covariation table corresponding to each subpopulation, and only then combined these two values to create a causal strength estimate for the whole population. If participants computed causal strength using ΔP, this alternative procedure would confirm equal strength across moderator conditions. However, if their computation of causal strength is best described by causal power, this alternative procedure would yield lower estimates for causal strength in the moderated condition, supporting an alternative explanation for the pattern observed in Experiment 1. However, this alternative would also predict an interaction between frequency and moderator in Experiment 2 (see Appendix S5 for more details). That such an interaction was not observed speaks against this account of the effects of moderation on causal judgments.

A third possibility is that our causal and explanatory judgments reflect assessments of either "simple" or "conjunctive" causal power, as defined by Novick and Cheng (2004). Again, we can evaluate this possibility with the data from Experiment 2. If participants interpreted our questions as requests to evaluate the simple causal power of the putative cause variable—that is, its capacity to produce the effect *in the absence* of the potentially interacting additional variable (the moderator)—then the data we provided to participants would produce lower simple power ratings in the moderated than non-moderated conditions, even if participants were not responsive to stability per se. In Appendix S1, we show that this account makes predictions contradicted by our data.

A fourth possibility is that participants instead interpreted our questions as corresponding to some weighted combination of simple and conjunctive causal powers. In this case, though, we should again have observed an interaction between frequency and moderation, with an upward trend in ratings in the moderated condition and no such trend in the non-

moderated condition (see Appendix S1 for more detail). Once again, this was not observed.

Our findings therefore provide strong support for the idea that causal and explanatory judgments are sensitive to stability as such, and suggest that these ratings are not a straightforward function of (existing) measures of causal strength.

4. Experiment 3

Our goals in Experiment 3 were threefold. First, while the previous experiments provided strong evidence for an effect of stability on causal and explanatory judgments across a variety of contexts and measures, they had two important limitations. For one thing, these experiments used artificial stimuli involving alien creatures and kinds. In addition, participants were presented with full numerical information (in the form of covariation tables), which is rarely if ever available in real life. Thus, our first goal in Experiment 3 was to replicate the effects of stability with more naturalistic stimuli. To this end, we presented participants with a hypothetical but realistic scenario describing the results from a study of the relationship between daily intake of folate (a kind of B vitamin) and increased bone density. As before, participants were presented with evidence of a causal relationship between these two factors (although not in the form of covariation tables), and we varied whether the relationship was moderated or non-moderated; the moderating variable was whether the participants of the hypothetical study carried variant A or variant B of a certain gene.

Second, the stimuli in Experiment 3 were designed to prevent participants from a potential misinterpretation of the moderator effect in Experiments 1–2. While the data presented to participants in all experiments suggested that the moderator was not an independent and sufficient cause of the effect (in fact, the data were inconsistent with this interpretation), there is a chance that participants could misconstrue it as such, introducing an alternative, competing cause. In addition to addressing this concern in the experiment reported in Appendix S4, we designed the cover story in Experiment 3 to make this misinterpretation unlikely: Carrying a particular gene variant is unlikely to cause a significant increase in bone density over the year targeted in the described study. Thus, any difference in causal ratings of folate between moderated and non-moderated conditions in this study would not support an alternative explanation of our findings in terms of some perceived competition between two potential direct and sufficient causes.

Our final goal was to explore whether effects of stability extend beyond causal and explanatory judgments to influence decisions and actions. The practical implications of stability have already been highlighted in the philosophical literature. Woodward (2010, 2016), for example, emphasizes that stable causal relationships are valuable insofar as they better subserve the goals of prediction and control. While the relationships between stability and practical reasoning could be explicated in various ways, in Experiment 3, we tested a particularly strong hypothesis about the practical effects of stability: the hypothesis that stable causal relationships are regarded as better guides to action than

unstable causal relationships, *even when the expected utility of the two courses of action is held fixed*. To test this hypothesis, we asked participants whether they themselves would be willing to take folate supplementation in light of the information with which they were presented. Because causal strength was held fixed across the moderated and non-moderated conditions and because participants were unaware of which gene variant they possessed, the expected utility of taking folate was the same in both conditions. To explore a wider range of possible practical effects of stability we also asked participants to answer additional questions, described below.

4.1. Method

4.1.1. Participants
A total of 201 participants (excluding an additional 119 participants who failed a memory check[11]) were recruited on Amazon Mechanical Turk in exchange for $0.60.

4.1.2. Materials, design, and procedure
The design and procedure were similar to those of Experiments 1 and 2 (with exceptions noted below), but the materials were different: Instead of stories about novel natural kinds on a fictional planet, participants read a realistic report about a hypothetical medical experiment. To increase the realism of the scenario, the results were not presented in the form of full covariation tables (which are rarely available in real-life situations).

Below is the text that participants read (non-moderated condition wording shown in square brackets):

Bone density is an important health factor. Fragile bones increase the risk of fractures, which is associated with other severe health complications, especially at later stages of life.

Researchers at the University of Manitoba have recently conducted a study investigating the potential effects of folate (a kind of B vitamin) on bone density. To test whether folate might improve density of bone tissues, the researchers recruited 2,000 participants and randomly assigned each of them to either a *treatment* or *control* condition. The 1,000 participants assigned to the treatment group were instructed to take a daily folate supplement for 1 year. The 1,000 participants assigned to the control group were given a placebo pill (a pill that unbeknownst to them did not contain any folate) to be taken daily for 1 year.

The researchers measured bone density for each participant both at the beginning and the end of the experiment. At the beginning of the experiment, there were no differences in average bone density between the two groups. But the researchers found that *after the experiment was over, average bone density in the treatment group was twice as high as in the control group.*

As a follow-up, the researchers decided to study whether genetic factors might make a difference to the effects of folate on bone density. They decided to focus specifically on the EXT1 gene, as this gene is known to play an important role in many different biological processes.

There are two equally frequent variants of the EXT1 gene: the A-variant and the B-variant. To investigate whether these variants of EXT1 might make a difference to the effect of folate, the researchers analyzed the genome of each participant in order to determine which variant of the gene he or she possessed. As expected, they found that roughly half of the participants in the experiment had variant A and roughly half had variant B. Then they looked at the results of the experiment for participants with gene variant A and for participants with gene variant B.

They found that the results of the experiment varied among the two groups: *while the average bone density of treatment participants with gene variant A was much higher than in the control group, the average bone density of treatment participants with gene variant B was the same as in the control group. That is, while daily folate intake was associated with increased bone density for participants with variant A, it was* not *associated with increased bone density among participants with variant B.* [The researchers found that there was no difference between the two variants: *the average bone density of treatment participants with gene variant A was twice as high as in the control group, and the average bone density of treatment participants with gene variant B was also twice as high as in the control group.*]

Thus, as before, we varied whether the results of the fictional experiment indicated the presence of a moderated or non-moderated relationship. Because participants in both conditions were told that average bone density was twice as high in the treatment group as in the control group, the average causal strength of the causal relationship between folate and bone density was matched across the non-moderated and the moderated conditions.

After reading this information, participants were asked to answer two sets of questions, presented in random order. In one block, participants were asked to evaluate unqualified claims about the causal relationship between folate supplementation and bone density, where each claim was presented at either the type or the token level (see Table 3). (Because previous experiments did not reveal significant differences between causal and explanatory claims, we dropped explanatory claims from Experiment 3.)

In another block we asked a series of questions concerning possible actions participants might take; we did so to investigate whether stability can influence real-life decisions and actions. The measure of main theoretical interest targeted the relationships between stability and intervention. Participants were asked: "Based on what you have read, how likely are you to consider folate supplementation for yourself?"; participants responded on a scale from 1 (not at all likely) to 7 (very likely). Note that because the average causal strength of folate supplementation on increased bone density is the same in the moderated and non-moderated conditions and participants do not know which variant of the gene

they have, the expected utility of taking folate supplementation is the same in both conditions. Thus, a significantly higher intervention rating in the non-moderated than the moderated condition would indicate that stability considerations can influence intentions to intervene, and it would additionally suggest that participants' dislike of unstable relationships cannot be brushed off as a mere pragmatic penalty for an underinformative speaker failing to mention a relevant factor when describing a relationship.

To explore a range of possible effects of stability beyond judgments of intervention, we included two additional measures. The "seeking information" question asked: "Based on what you have read, would you be interested in receiving a free brochure on folate supplements?" (1 not at all interested—7 very interested). Finally, the "resource investment likelihood" question said: "The Mayo clinic will soon be conducting a clinical trial where participants will be provided with free monthly supplies of two leading brand folate supplements. The trial will take about 1 hour of your time. Based on what you have read, how likely would you be to sign up for the trial?" (1 not at all likely—7 very likely).

In addition to the main set of questions, participants answered two four-option, forced-choice memory check questions. First, right after reading the description of the hypothetical study, they were asked to select a statement correctly describing what the researchers found when they looked at the groups with gene variant A and B separately. The correct answer differed depending on the moderator condition. For the non-moderated condition, the correct answer read: "Daily folate intake was associated with equal amounts of increased bone density for participants with gene variant A and gene variant B." For the moderated condition, the correct answer read: "Daily folate intake was associated with improvements in bone density only for participants with gene variant A, but there was no improvement for participants with gene variant B." The two filler statements were inaccurate for both conditions.

Second, at the very end of the survey they were asked what the researchers in the hypothetical study did after dividing study participants in two groups, with the correct answer being that "they gave folate supplements to one group, but not the other." Participants who answered one or both questions incorrectly were excluded from analyses.

Table 3
Causal judgments used in Experiment 3, as a function of target (type vs. token)

Type	How much do you agree with the following causal statement? *Folate supplementation causes an increase in bone density*
Token	Marina was a participant in the University of Manitoba study. She was assigned to the treatment group, and thus took daily folate supplements for one year. At the end of the experiment, Marina's bones have increased in density. Marina's records were lost, so nobody knows whether she has variant A or B of the EXT1 gene. How much do you agree with the following causal statement? *Folate supplementation caused Marina's bones to become denser.*

Thus, Experiment 3 had a 2 moderator (moderated vs. non-moderated relationship) × 2 target (type vs. token) between-subjects design. (Since target was not a variable of central theoretical interest, we report effects involving this variable only when they were significant, as we did in the previous experiment.) The dependent variables were agreement with causal claims measured on a scale of 1 (strongly disagree) to 7 (strongly agree) and action ratings measured on a scale of 1 (not at all likely/not at all interested) to 7 (very likely/very interested).

4.2. Results and discussion

4.2.1. Causal ratings

Our main question regarding causal ratings was whether the effects of stability on causal judgments observed in previous experiments still held in the context of the more realistic scenario used in the present experiment. An independent samples t-test on unqualified causal ratings as a function of moderator condition showed that participants were more likely to endorse the causal claim about the causal relationship between folate supplementation and bone density when the relationship was non-moderated than moderated, $t(199) = 3.64$, $p < .001$, Cohen's $d = .51$ (see Fig. 5a).[12]

4.2.2. Action ratings

To examine whether the stability of a causal relationship has an effect on participants' willingness to act on the relationship by intervening on the cause, even when the expected utility of the intervention is equated across stability conditions, we analyzed participants' answers to the question "Based on what you have read, how likely are you to consider folate supplementation for yourself?" Participants were significantly more likely to consider folate supplementation for themselves in the non-moderator than moderator condition, $t(199) = 2.93$, $p = .004$, Cohen's $d = 0.42$, even though the expected utility of doing so was the same in both conditions (see Fig. 5b).

For the remaining questions, the moderator manipulation did not affect ratings. Participants were not significantly more interested in receiving a free brochure about folate in the non-moderated than the moderated condition, $M_{non\text{-}mod} = 4.13$, $SD = 2.04$, $M_{mod} = 3.83$, $SD = 2.09$, $t(199) = 1.03$, $p = .304$, nor were they more willing to invest an hour of their time to participate in a clinical trial, $M_{non\text{-}mod} = 4.58$, $SD = 2.03$, $M_{mod} = 4.46$, $SD = 2.01$, $t(199) = 0.40$, $p = .692$ (see Fig. 5c and d).

In sum, we replicate the effect of stability on causal ratings observed in Experiments 1 and 2, demonstrating that stability considerations play an important role in "real-life" scenarios involving causal judgments. Additionally, we found that participants were more likely to intervene on a cause when the relationship was stable rather than unstable, even when the expected utility of such actions was held fixed across conditions. This suggests that stable causal relationships are regarded as better guides to action than unstable causal relationships.[13]
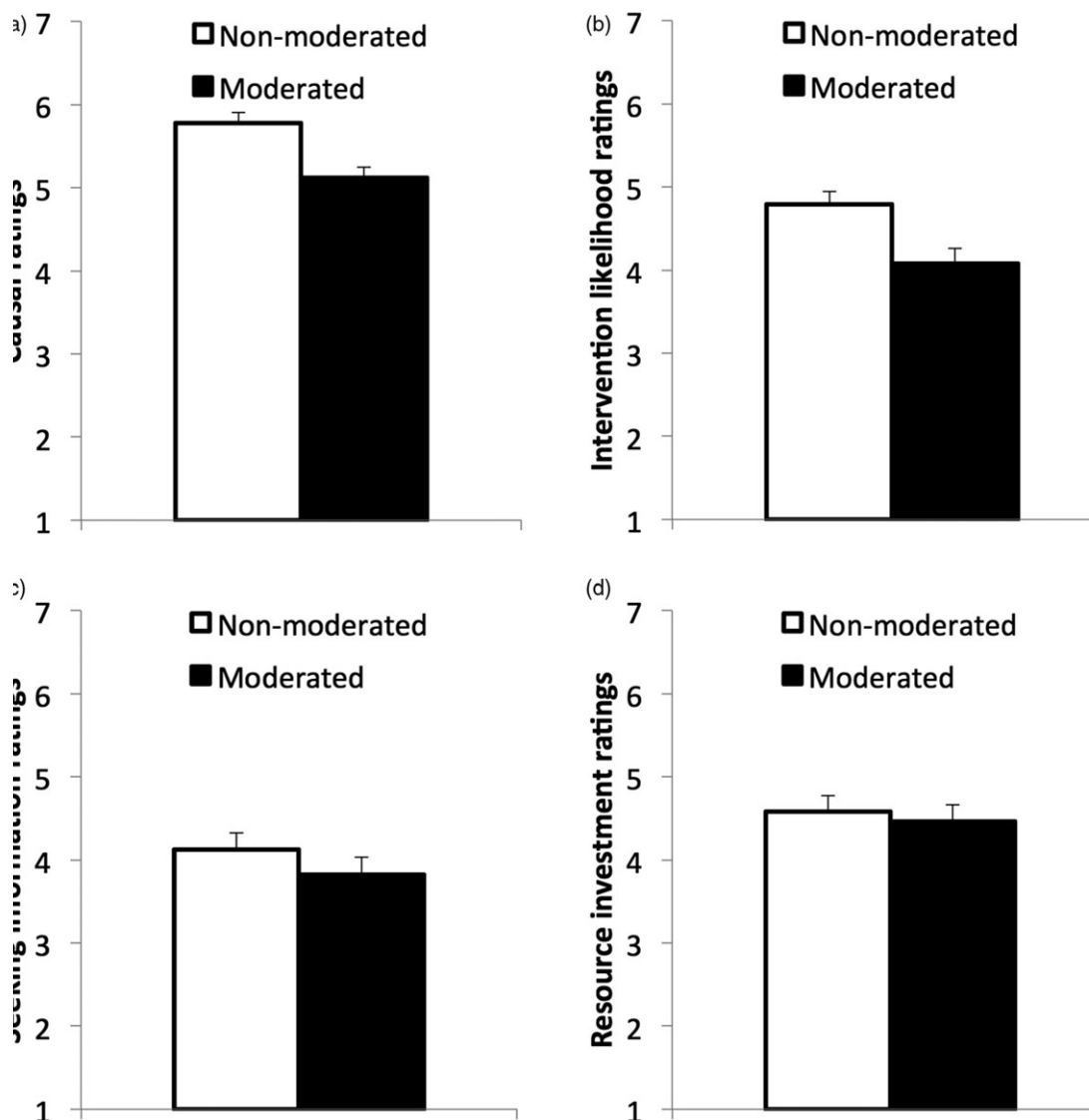
Fig. 5. The effect of moderator on causal ratings (a), intervention likelihood ratings (b), information seeking ratings (c), and resource investment ratings (d) in Experiment 3. Error bars correspond to 1 SEM.

5. General discussion

In three experiments, we document an important factor shaping people's assessments of causal and explanatory relationships over and above causal strength: the *stability* of the causal relationship—that is, the extent to which it holds across various possible circumstances. While philosophers of science have stressed the importance of stability in

scientific modeling and explanation, the role of stability in causal and explanatory judgments has not received adequate attention within psychology. The three reported experiments show that stability considerations play a consistent role in various contexts: People are more willing to endorse causal and explanatory claims involving stable causal relationships for statements at both type and token levels. The results of Experiment 3 additionally show that these effects hold when people reason about realistic scenarios, and that people may take stable causal relationships to be better guides to action than their unstable counterparts, even when the difference in stability does not translate into a difference in expected utility.

Our results rule out several alternative explanations for the effects we attribute to stability. First, the results of Experiment 2 suggest that the effect of stability is not reducible to actual scope (that is, unstable causal relationships are not penalized merely because they hold for a smaller actual proportion of the population). Second, our results suggest that effects of stability are not limited to one particular question or locution, since the effect was observed for both causal and explanatory claims, at the type and token levels, and for questions about causal structure and causal strength. Third, our results suggest that unstable claims were not penalized merely for pragmatic infelicity (i.e., for failing to specify a moderating circumstance when it was present), since the effect extended to an intervention judgment in Experiment 3. Finally, it is noteworthy that the effect of stability was observed even when participants learned about the relevant variability in relationship strength through descriptive summary statements (as in Experiment 3) rather than having to extract this information from covariation tables. This indicates that stability can be assessed from multiple kinds of evidence, and poses an additional challenge to attempts to account for the entirety of our findings by tweaks in calculations over covariation tables. Overall, our results suggest a robust effect of stability across a variety of claim types and data presentation formats, with causal and explanatory generalizations penalized when they apply in a limited range of circumstances.

6. Relationship to prior research

Stability is an aspect of a causal relationship that is distinct from traditional measures of causal strength, such as ΔP (Allan, 1980) or power PC (Cheng, 1997). These measures track one aspect of causal relationships: their average strength in a population. However, they do not capture another important aspect that matters for causal assessment (and in particular for generalization), namely the extent to which the relationship holds in a range of plausibly-occurring background circumstances. An important question, though, concerns the relationship between stability in causal generalizations and measures of causal strength. In particular, can the effect of stability be expressed in terms of these existing measures?

The discussion of Experiment 2, and Appendix S1, outline some reasons why the effect of stability on causal generalization is unlikely to arise as a function of causal strength estimates over simple or conjunctive causal influences. That said, the effects we

report have close connections to people's capacity to identify unstable relationships from observed evidence. In fact, prior work has recognized the identification of stable, or "invariant," relationships as an important objective of causal learning and demonstrated that people can infer the presence of background factors interacting with a causal relationship based on differences in covariation across contexts (Liljeholm & Cheng, 2007; see also Clifford & Cheng, 2000; as cited in Novick & Cheng, 2004; White, 1998, for additional evidence that people can assess conjunctive causation from covariation data). Taken together, this work demonstrates that people are able to track the kind of evidence relevant to assessments of stability; what our experiments show is that stability will in turn affect the endorsement of general causal claims, as well as ratings of causal structure and strength.

An important direction for future research is to better understand how people arrive at assessments of stability more generally. The results of Experiment 3, for example, cannot readily be explained by analyses over covariation tables. Instead, we suspect that people rely on a host of cues to stability and generalization, with some kinds of instability posing a greater threat to the felicity of a causal generalization. We expect that research on the evaluation of generic claims (e.g., Cimpian & Markman, 2011; Cohen, 1999; Gelman, Star, & Flukes, 2002; Goldin-Meadow, Gelman, & Mylander, 2005; Leslie, 2014; Pelletier, 2009; Prasada & Dillingham, 2006, 2009; Prasada, Khemlani, Leslie, & Glucksberg, 2013; Tessler & Goodman, 2016) may be especially helpful in illuminating the relationship between observations and causal generalizations, which often take generic form.

Other work on causal reasoning and representation has aimed to identify necessary and sufficient conditions for causal ascription, or to differentiate relationships of causation from those of "allowing," "enabling," or "preventing" (e.g., Goldvarg & Johnson-Laird, 2001; Khemlani, Barbey, & Johnson-Laird, 2014; Sloman, Barbey, & Hotaling, 2009; Wolff & Barbey, 2015). These notions could have some relationship with stability. For instance, people may be more inclined to reject a causal ascription when it highlights an unstable relationship, or to include enabling factors in a causal claim when doing so will substantially boost its stability. So while this prior work has not itself investigated stability, it provides methods and phenomena that could be of value in pursuing a better understanding of how and why stability influences causal generalizations.

As mentioned in the introduction, a handful of prior studies provide indirect evidence that stability considerations can influence causal generalizations. Beyond the findings from Lombrozo (2010), there is evidence that people are less inclined to regard an agent as a cause of a bad outcome when a third-party intentionally controlled the agent (Phillips & Shaw, 2015; Murray & Lombrozo, 2017). A possible explanation suggested by Murray and Lombrozo (2017) is that the dependence of the outcome on the agent is very sensitive to the third-party's intentions, and in that respect fairly unstable. Along with the present findings, this work supports the *exportability* theory of explanation (Lombrozo & Carey, 2006) and causal ascription (Lombrozo, 2010), according to which a central function of explanations and causal ascriptions is to pick out patterns of dependence that are exportable in the sense that they support future predictions and interventions. If this is

correct, we should expect explanatory and causal ratings to favor more stable relationships: by being insensitive to variations in background circumstances, a stable causal relationship typically provides more opportunities for effective prediction and intervention.[14]

That said, there are two distinct senses in which causal generalizations might be stable, and thus exportable. Blanchard, Vasilyeva, and Lombrozo (2017) argue for a distinction between two notions of stability: breadth and guidance. Breadth reflects the range of background circumstances in which a causal generalization holds: a broader generalization holds in every circumstance where a narrower generalization holds, plus in additional circumstances. Guidance, as the name suggests, reflects the amount of support a causal generalization provides for developing informed expectations about whether the causal relationship will or will not hold in particular circumstances. Both breadth and guidance contribute to the explanatory value of a causal generalization, but they do so by maximizing different benefits: the former favors more inclusive generalizations (covering more cases), while the latter favors more exclusive generalizations (correctly excluding cases where the relationship is unlikely to hold).

To illustrate this distinction, consider again the causal relationship between eating yonas and getting sore antennas when it holds only in one background circumstance (e.g., if the zelmo's diet also contains salty water). The causal generalization "eating yonas causes sore antennas" is unstable in two senses: It holds in a narrow range of circumstances and it provides poor guidance for generalization to novel cases. One way to alleviate instability is to explicitly build the background circumstance into the relationship: "*For zelmos who drink salty water*, eating yonas causes sore antennas." This qualified claim seems better than the bare claim not because it applies to a wider range of possible circumstances *per se*, but because it is more "guiding": By flagging the circumstance under which the relationship holds, it provides a better sense of *when* the relevant causal relationship can be used for prediction and control, and it is therefore exportable in the sense that it contains conditions for application, whether or not those conditions hold widely. Indeed, it is important to note that on Woodward's account, evaluations of stability are not restricted to single causes (see Woodward, 2010, p. 289): A conjunction of causes can support a stable causal generalization, and a causal generalization that specifies both relevant values of the cause and moderator variables can maximize stability in Woodward's sense, and guidance in ours.

Before turning to further open questions, it is worth considering two recent papers in greater detail, as both explore ideas closely related to stability as we define it here. Using a very different paradigm involving collisions between physical objects, Gerstenberg et al. (2012) found that the "robustness" of an outcome (i.e., whether a ball that was hit by another ball *clearly* or *barely* went through a gate) did not affect "cause" versus "prevent" judgments (e.g., "A caused B to go [prevented B from going] through the hole"). However, robustness did predict choices between descriptors of causal relationships ("caused," "prevented," "almost caused/prevented," "helped [to prevent]"), and it had some effect on the responsibility assigned to potentially competing causes in complex causal structures, including causal chains (Gerstenberg et al., 2015). This work provides additional evidence that considerations relevant to stability can influence psychological

judgments, but the differences across the paradigms makes a direct comparison to our task quite difficult. In particular, Gerstenberg and colleagues' task was designed to measure judgments concerning counterfactuals and the choice of causal descriptors, not to equate causal strength or assess generalizations over many actual instances.

In a different line of recent work, Nagel and Stephan (2016) tested the hypothesis that a cause A will be regarded as a better explanation of some outcome C when the mediating mechanism B is itself "insensitive" in the sense that it holds beyond the specific case under evaluation. Participants evaluated claims about how crucial a cause was for producing an outcome in a particular fish or machine, before and after learning that the cause was (or was not) reliably associated with the outcome in other similar fish or machines. As predicted, participants more strongly endorsed the claim that A is "crucial for" C when the A->B->C chain was observed in cases beyond that under evaluation. While our findings are consistent with those of Nagel and Stephan, the specifics of their method make it less suitable for evaluating our proposal, given how we define stability here. Namely, the judgments elicited were for causal necessity rather than endorsement of causal and explanatory generalizations, and the stability manipulation at the population level (all observed fish or machines) was confounded with causal strength. Thus, while Nagel and Stephan's study makes an independently valuable contribution, it does not establish what we aim to demonstrate with our own studies: that general causal and explanatory claims are influenced by stability, above and beyond causal strength.

7. Limitations and open questions

Our findings suggest several additional directions for future research. First, while our results document the effect of stability, they underspecify the underlying mechanism. For example, stability considerations could influence people's perceptions of the causal strength of the relevant relationships, and/or they could influence downstream assessments of causal and explanatory generalizations. Our results from Experiment 3, which did not involve computing causal strength from covariation tables, suggest that effects of stability can manifest after the strengths of the relevant causal relationships have been computed, but many possibilities remain. Describing the overall mechanism in further detail is an important avenue for future research.

Second, how does stability connect with issues of simplicity in causal representation? As Woodward (2016) notes, causal structures involving stable relationships can be represented with sparse causal graphs, whereas unstable relationships complicate the task of causal representation (see also Powell, Merrick, Lu, & Holyoak, 2016). It could be that a preference for more stable claims is partially a consequence of other representational preferences.

Third, our manipulation of stability was concerned with the number of circumstances in which the target relationship held. However, Woodward (2006, 2010) also emphasizes a role for the normality and importance of background circumstances over which stability is calculated. Indeed, extensive evidence shows that causal claims are

influenced by considerations of normality (e.g., Hitchcock & Knobe, 2009; Icard & Knobe, 2016). A natural question for further research is how the number, normality, and importance of background circumstances jointly influence causal and explanatory judgments. The results of Experiment 2, it is worth noting, suggest that variations in the frequency of the background circumstance in the population does not affect the perceived stability of a causal relationship, at least when this frequency does not reach extreme values.

Fourth, can stability account for intransitivity in causal chains? For instance, it is reasonable to say that sex causes pregnancy, and that pregnancy causes nausea, but it seems less reasonable to say that sex causes nausea. Johnson and Ahn (2015) show that causal chains with equally strong intermediate links may nevertheless differ in transitivity, and argue that some causal relations must be represented as "causal islands" rather than coherent networks. Could stability help explain what makes some causal relations behave as causal islands (regardless of the nature of the representation)? For example, intransitivity could arise if the component links are evaluated with respect to different sets of moderators, and/or there is little overlap between the subsets of background circumstances for which the component relationships hold. Relatedly, Nagel and Stephan (2015) show that when A causes B, which in turn causes C, people are less willing to regard A as an appropriate explanation of C when the mediating factor is an abnormal intentional action (as opposed to a biological mechanism). As they point out, a possible explanation of this fact is that causal relationships mediated by abnormal intentional actions are perceived as less stable.

Finally, *why* do stability considerations influence causal and explanatory generalizations? Both philosophers and psychologists have suggested that stable relationships better support predictions and interventions, and our own findings from Experiment 3 point in this direction. Nonetheless, many questions remain about whether and how these generalizations influence subsequent judgments and behaviors, and when this influence is beneficial.

Acknowledgments

Notes

1. Note that "background circumstance" here has a meaning different from the one it has in the literature on "the problem of causal selection," that is, the problem of explaining why for instance we regard the lighting of the match as a "real cause" of the fire and the presence of oxygen as a mere "background circumstance." As

Woodward uses the term, a "background circumstance" (relative to a causal relationship $X \rightarrow Y$) is any circumstance that is not explicitly represented by X or Y, and such a circumstance may well be a "real cause" of Y rather than a "mere background condition." Note also that there may be differences between effects of stability when it is defined with respect to *background circumstances* as defined here versus with respect to the *manner* in which the cause occurs (Lombrozo, 2010), and/or to the *status of intermediate causes* in a causal chain (e.g., as in Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2015); investigation of these differences is beyond the scope of this paper.

2. We thank Jonas Nagel for encouraging us to clarify the relationship between stability and the components of an ANOVA.

3. That is, the non-moderated condition strength never exceeded the moderated condition strength, making it harder to demonstrate the effect we predicted. This also held for other metrics of causal strength computed over covariation tables (pooling across moderator variable values), for example, causal power as formulated in Cheng (1997) (see Cheng & Holyoak, 1995, on assessment of a causal contingency pooling across multiple focal sets).

4. In all experiments, an additional group of participants ($N_{Exp1} = 86$; $N_{Exp2} = 209$; $N_{Exp3} = 42$, excluding 30, 42, and 23 participants, correspondingly, who failed memory checks) evaluated qualified type-level causal (and, in Experiments 1 and 2, explanatory) claims that specified the subgroup defined by the moderator variable, for example, *For zelmos who drank salty water, eating yonas causes their antennas to become sore* (one rating for the "high-moderator" subgroup, e.g., salty water, where the moderator variable was set to a value at which the moderated "yona→ sore antenna" relationship is strong, and one rating for the "low-moderator" subgroup, e.g., fresh water). As expected, the ratings reflected the differences in the co-variation tables presented to participants across conditions: for moderated relationships, the ratings were higher for the high-moderator subgroup than the low-moderator subgroup, $p_{Exp1} < .001$, $p_{Exp2} < .001$, $p_{Exp3} < .001$; when a relationship was not moderated, that is, the split tables showed the same co-variation strength, participants' ratings did not differ between the two subgroups, $p_{Exp1} = .130$, $p_{Exp2} = .052$, $p_{Exp3} = .615$. Also in accordance with the provided data, across all three experiments, the ratings for the high-moderator subgroup were higher in the moderated condition than in the non-moderated condition, $p_{Exp1} = .018$, $p_{Exp2} = .004$, $p_{Exp3} < .001$, with the reversed pattern for the low-moderator subgroup, $p_{Exp1} < .001$, $p_{Exp2} < .001$, $p_{Exp3} < .001$. The moderator by subgroup interaction was significant in all experiments, all p's $< .001$. For the qualified group in Experiment 2, the co-variation strengths were controlled as described in the Appendix S6. In the main text, we focus on results from unqualified claims only as the ratings relevant for evaluating our hypothesis. See Appendix S1 for some additional analyses involving qualified claims.

5. Main effect of judgment $F(1, 178) = 0.03$, $p = .861$; main effect of target $F(1, 178) = .02$, $p = .894$; moderator by judgment interaction $F(1, 178) = 1.58$,

$p = .211$; moderator by target interaction $F(1, 178) = .16$, $p = .686$; judgment by target interaction $F(1, 178) = .27$, $p = .604$; three-way interaction $F(1, 178) = .01$, $p = .931$.

6. The constraint of equating average strength across the moderated and non-moderated relationships in all frequency conditions, combined with the constraint of keeping causal strength near zero in one subgroup (e.g., fresh water), entailed that the strength of the relationship in the subgroup with a causal association inevitably had to vary ($\Delta P = 0.97$, $\Delta P = 0.61$, and $\Delta P = 0.44$ for low, medium, and high frequency, respectively). This distribution of strength levels predicts a pattern of ratings in the moderated condition opposite to the pattern we should observe if the moderator effect is driven by variations in actual scope. If both effects take place, and they happen to be exactly equal in magnitude, they could cancel each other, complicating the interpretation of the results. See Appendix S6 for the results of an additional study ruling out this possibility.

7. An additional 2 moderator (moderated relationship, non-moderated relationship) × 2 judgment (causal, explanatory) × 2 target (type, token) × 3 moderator frequency (low, medium, high) mixed ANOVA on main causal and explanation ratings revealed that type ratings ($M = 4.63$) were higher than token ratings ($M = 4.26$, $F(1, 381) = 8.97$, $p = .003$, $\eta_p^2 = 0.023$), but there were no other main effects nor interactions (all $ps \geq .154$: effect of judgment type $F(1, 381) = 1.76$, $p = .186$; effect of frequency $F(2, 381) = 1.60$, $p = .203$; moderator x judgment $F(1, 381) = .02$, $p = .877$; moderator × target $F(1, 381) = 0.25$, $p = .620$; judgment × target $F(1, 381) = 1.46$, $p = .228$; judgment × frequency $F(2, 381) = 1.14$, $p = .322$; target × frequency $F(2, 381) = .78$, $p = .461$; judgment × target × frequency $F(2, 381) = .05$, $p = .951$; moderator × judgment × target $F(1, 381) < .01$, $p = .953$; moderator x judgment x frequency $F(2, 381) = .66$, $p = .519$; moderator × target × frequency $F(2, 381) = .89$, $p = .414$; moderator × judgment × target × frequency $F(2, 381) = .70$, $p = .496$).

8. The same analysis adding target as a between-subject factor replicated these results, but also produced an unpredicted interaction between frequency and target, $F(2, 387) = 3.35$, $p = .036$, $\eta_p^2 = 0.017$, driven by a drop in token ratings in the high-frequency condition. As a result, token ratings were lower than type in the high-frequency condition, $p = .010$, but not in other conditions, $ps \geq .293$; as another way of describing the same pattern, token ratings in the high-frequency conditions were lower than token ratings at other frequency levels, p's $\leq .018$, with no parallel effects in the type condition, all p's $\geq .102$. This interaction is not relevant for evaluating our main hypothesis about the effect of stability on causal judgments. No other effects were significant (all $ps \geq .170$).

9. Additionally, in the analysis including target and judgment factors, the interaction between frequency and target approached significance, $F(2, 386) = 2.66$, $p = .072$; similarly to the structure ratings, it appeared to be driven by a marginal drop for token ratings from the medium- to high-frequency condition ($p = .054$), resulting in lower token ratings than type ratings in the high-frequency condition ($p = .007$). This marginal interaction is irrelevant for evaluating our hypothesis.

10. Participants gave higher structure and strength ratings to the moderator variable→ effect relationship in the moderated (M_{stru} = 5.13; M_{stre} = 4.75) than non-moderated condition (M_{stru} = 2.91; M_{stre} = 2.34; $F_{stru}(1, 390)$ = 296.68, $p < .001$, $\eta_p^2 = 0.432$; F_{stre} = (1, 390) = 296.43, $p < .001$, $\eta_p^2 = 0.432$). The ratings were also higher in the high-frequency condition (M_{stru} = 4.35; M_{stre} = 4.00) than in low (M_{stru} = 3.85; M_{stre} = 3.32) and medium (M_{stru} = 3.87, M_{stre} = 3.32; Tukey HSD p's $\leq .005$), which did not differ from each other ($p \geq .991$; $F_{stru}(2, 390)$ = 6.75, $p = .001$, $\eta_p^2 = 0.033$; $F_{stre}(2, 390)$ = 10.24, $p < .001$, $\eta_p^2 = 0.050$). We did not predict this effect, but it is possible that participants' ratings were driven by the fact that in the high-frequency condition, the sheer number of (e.g.,) zelmos who drank salty water and developed sore antennas was higher than in the other frequency conditions. The moderator x frequency interaction was not significant on the structure measure, $F(2, 390)$ = 1.56, $p = .212$, but it did reach significance on the strength measure: Frequency only affected ratings in the non-moderated condition (higher ratings in high frequency than low and medium frequency, ps $< .001$, which did not differ from each other, $p = .854$), but the ratings in the moderated condition did not vary across frequency levels (all ps $\geq .149$). In both moderated and non-moderated conditions, the structure ratings of the moderator variable-effect relationships were significantly above the lowest scale endpoint of one (M_{mod} = 5.13, $t(392)$ = 45.95, $p < .001$; $M_{non\text{-}mod}$ = 2.91, $t(392)$ = 20.97, $p < .001$), and strength ratings were significantly above the lowest scale point of zero (M_{mod} = 4.75, $t(392)$ = 48.57, $p < .001$, $M_{non\text{-}mod}$ = 2.34, $t(392)$ = 22.36, $p < .001$).

11. Experiment 3 had higher exclusion rate than the previous experiments; we suspect this is due to a lack of visual aids and a longer scenario text than in Experiments 1 and 2 (we are grateful to an anonymous reviewer for suggesting this).

12. An ANOVA on the same data including target as a between-subjects factor showed the same effect, plus a significant main effect of target: Token ratings (M = 5.73) were higher than type ratings (M = 5.20), $F(1, 197)$ = 9.02, $p = .003$, $\eta_p^2 = 0.044$, but this effect did not interact with the moderator condition, $F(1, 197)$ = .26, $p = .614$.

13. The findings of Experiment 3 are reminiscent of Ellsberg's paradox (Ellsberg, 1961), where ambiguity intolerance leads to violations of expected utility theory. One of Ellsberg's cases involves two urns: The first contains 50 red balls and 50 black balls, while the second contains 100 balls with red and black proportions left unspecified. Ellsberg shows that in such a situation, people display a preference for bets on draws from the first urn to draws from the second urn, in a way that violates the axioms of expected utility theory. However, our findings may be only superficially related to Ellsberg's paradox. The paradox is a manifestation of a preference for risk over uncertainty (in the technical sense of these terms). In our experiment, however, participants in the moderated condition do not face more uncertainty than participants in the non-moderated condition; in particular, in both conditions participants are provided with explicit information about the

proportions of people with variants A and B (namely 50/50). We are grateful to Tobias Gerstenberg for pointing out this connection.

14. We do not mean to suggest that stability is all that matters for prediction and intervention: causal strength (and probably other factors) matter as well. For instance, an irreducibly probabilistic *C-E* relationship such that *C* has a 5% chance of producing *E* in all contexts is very stable but does not support prediction very well. (We are grateful to Jonas Nagel for bringing this example to our attention.) Nevertheless, when compared to an equally weak (on average) but unstable relationship, the latter will produce greater prediction error in estimates for a rate of occurrence of effect *E* given *C* in a *sample* drawn under *unknown but uniform* background circumstances. Thus, on average, stable relationships still provide a better basis for prediction.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147–149. https://doi.org/10.3758/BF03334492.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2017). Stability, breadth and guidance. *Philosophical Studies*. https://doi.org/10.1007/s11098-017-0958-6

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

Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 227–253). Cambridge, MA: MIT Press.

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271–302). MIT Press, Cambridge, MA.

Cheng, P. W., Liljeholm, M., & Sandhofer, C. M. (2013). Logical consistency and objectivity in causal learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 2034–2039). Cognitive Science Society Austin, Austin, TX.

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492. https://doi.org/10.1111/j.1467-8624.2010.01525.x.

Cohen, A. (1999). Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, *22*(3), 221–253. https://doi.org/10.1023/A:1005497727784.

Gelman, S. A., Star, J. R., & Flukes, J. (2002). Children's use of generics in inductive inferences. *Journal of Cognition and Development*, *3*(2), 179–199. https://doi.org/10.1207/S15327647JCD0302_3.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles & R. P. Cooper, (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 378–383. Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, & P. P. Maglio, (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX. Cognitive Science Society.

Goldin-Meadow, S., Gelman, S. A., & Mylander, C. (2005). Expressing generic concepts with and without a language model. *Cognition*, *96*(2), 109–126. https://doi.org/10.1016/j.cognition.2004.07.003.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565–610. https://doi.org/10.1207/s15516709cog2504_3.

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

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612. https://doi.org/10.5840/jphil20091061128.

Icard, T. F., & Knobe, J. (2016). Causality, normality and sampling propensity. In A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 800–805). Austin, TX: Cognitive Science Society.

Johnson, S. G., & Ahn, W. K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, *39*(7), 1468–1503. https://doi.org/10.1111/cogs.12213.

Kendler, K. (2005). A gene for…: The nature of gene action in psychiatric disorders. *American Journal of Psychiatry*, *162*, 1243–1252. https://doi.org/10.1176/appi.ajp.162.7.1243.

Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience*, *8*, 849. https://doi.org/10.3389/fnhum.2014.00849.

Leslie, S.-J. (2014). Carving Up the Social World with Generics. In T. Lombrozo, J. Knobe & S. Nichols (Eds.), *Oxford studies in experimental philosophy*, (pp. 208–232). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198718765.003.0009

Lewis, D. (1986). Postscript C to "Causation": Insensitive causation. *Philosophical Papers*. Vol. 2 (pp. 184–188). Oxford, UK: Oxford University Press.

Liljeholm, M., & Cheng, P. (2007). Coherent generalization across contexts. *Psychological Science*, *18*, 1014–1021. https://doi.org/10.1111/j.1467-9280.2007.02017.x.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009.

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attribution of causation, free will, and moral responsibility. *Cognitive Science*, *41*(2), 447–481. https://doi.org/10.1111/cogs.12338.

Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1691–1696). Austin, TX: Cognitive Science Society.

Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806–812). Austin, TX: Cognitive Science Society.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485. https://doi.org/10.1037/0033-295X.111.2.455.

Pelletier, F. J. (Ed.). (2009). *Kinds, things, and stuff: Mass terms and generics*. Oxford University Press; Oxford, UK. https://doi.org/10.1093/acprof:oso/9780195382891.001.0001

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science*, *39*(6), 1320–1347. https://doi.org/10.1111/cogs.12194.

Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive psychology*, *86*, 62–86. https://doi.org/10.1016/j.cogpsych.2016.02.001.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112. https://doi.org/10.1016/j.cognition.2005.01.003.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, *33*(3), 401–448. https://doi.org/10.1111/j.1551-6709.2009.01018.x.

Prasada, S., Khemlani, S., Leslie, S. J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–422. https://doi.org/10.1016/j.cognition.2012.11.010. doi: 10.1016/j.cognition.2012.11.010.

Sloman, S. (2005). Causal models: How people think about the world and its alternatives. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195183115.001.0001.

Sloman, S., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, *33*(1), 21–50. https://doi.org/10.1111/j.1551-6709.2008.01002.x.

Sloman, S., & Lagnado, D. A. (2003). Causal invariance in reasoning and learning. *Psychology of Learning and Motivation*, *44*, 287–325. https://doi.org/10.1016/S0079-7421(03)44009-7.

Spellman, B. A. (1996a). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, *7*(6), 337–342. https://doi.org/10.1111/j.1467-9280.1996.tb00385.x.

Spellman, B. A. (1996b). Conditionalizing causality. *Psychology of Learning and Motivation*, *34*, 167–206. https://doi.org/10.1016/S0079-7421(08)60561-7.

Tessler, M. H., & Goodman, N. D. (2016). A pragmatic theory of generic language.

Tessler, M. H., & Goodman, N. D. (2016). Communicating generalizations about events. In A., Papafragou, D., Grodner, D., Mirman, & J. C., Trueswell. (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

White, P. A. (1998). Causal judgement: Use of different types of contingency information as confirmatory and disconfirmatory. *European Journal of Cognitive Psychology*, *10*, 131–170. https://doi.org/10.1080/713752269.

Wolff, P., & Barbey, A. K. (2015). Causal reasoning with forces. *Frontiers in Human Neuroscience*, *9*, 1–21. https://doi.org/10.3389/fnhum.2015.0000

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review*, *115*, 1–50. https://doi.org/10.1215/00318108-115-1-1.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*, 287–318. https://doi.org/10.1007/s10539-010-9200-z.

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

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1 Relationship to Power PC framework (Cheng, 1997).

Appendix S2 Experiment 1 sample trial.

Appendix S3 Experiment 2 sample trial.

Appendix S4 An additional experiment evaluating an alternative explanation for differences across moderated and unmoderated conditions.

Appendix S5 A clarification of two approaches to averaging causal strength across subpopulations.

Appendix S6 An additional experiment using an alternative approach to fixing causal strength (compared to Experiment 2).