Stable Causal Relationships are Better Causal Relationships

Nadya Vasilyeva (Vasilyeva@Berkeley.Edu)*,
Thomas Blanchard (Tblanchard@Berkeley.Edu)*,
Tania Lombrozo (Lombrozo@Berkeley.Edu)
Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA

*These two authors contributed equally to this work

Abstract

We report two experiments investigating whether people's judgments about causal relationships are sensitive to the robustness or *stability* of such relationships across a wide range of background circumstances. We demonstrate that people prefer stable causal relationships even when overall causal strength is held constant, and we show that this effect is unlikely to be driven by a causal generalization's actual scope of application. This documents a previously unacknowledged factor that shapes people's causal reasoning.

Keywords: stability, causality, explanation, background conditions, moderating variables

Consider the relationship between being sexually active and developing deep vein thrombosis. Is the former a *cause* of the latter?

In fact, sexual activity increases the probability that a woman will become pregnant, which in turn increases the probability that she will develop deep vein thrombosis. Thus, there is a sense in which being sexually active is causally relevant to whether one gets thrombosis. Yet this causal relationship seems different from that between, say, sexual activity and contracting herpes. When explaining why someone contracted herpes (but not why a woman has deep vein thrombosis), we're likely to cite sexual activity. Similarly, medical websites list herpes as a sexually transmitted disease, but do not list deep vein thrombosis as a risk associated with sexual activity. Why might this be?

The asymmetry between sex and thrombosis versus sex and herpes doesn't seem due to a difference in the strength of association between the relevant factors. Few people suffer from thrombosis as a result of sexual activity; but likewise, herpes is only contracted after sexual activity in some cases (namely, when both the partner is infected and transmission occurs). And while thrombosis can be caused by many other factors besides sexual activity, herpes can also be transmitted by non-sexual forms of physical contact.

Instead, these causal associations could differ in their robustness or *stability*. While sex can elevate the risk of thrombosis under very specific conditions (most notably, when the person is a woman who becomes pregnant), there are also plenty of circumstances under which there is simply *no* causal relationship between being sexually active and deep vein thrombosis, e.g., if the person is male, or sterile, or on blood thinners, and so on. By contrast, the causal relationship between sex and herpes is more robust insofar as under most circumstances – whatever your gender, your diet, your age, etc. – sexual activity puts you at greater risk of contracting herpes.

The extent to which a causal relationship is stable is not adequately captured by measures of causal strength that have dominated research on causal inference, such as ΔP (Allan, 1980) and power-PC (Cheng, 1997). The reason is that these measures track the *average* strength of a causal relationship in a population, whereas stability has to do with the extent to which the relationship holds *across* diverse segments of the population (or across various circumstances). True, very unstable causal relationships, such as the relationship between sexual activity and thrombosis, also tend to have low average strength. But two causal relationships can be equally strong on average and yet not be equally stable. By treating both relationships on a par, standard measures of causal strength ignore an important difference between them. For instance, these measures do not allow us to capture the fact that more stable relationships provide more far-reaching and reliable means for controlling their effects.

Stability is a well-known notion in the philosophy of science, where it has been introduced and discussed most extensively by Woodward (2006, 2010). In Woodward's framework, one starts with a very undemanding notion of causal relevance, on which X is causally relevant to Y just in case X causally influences Y in at least *some* circumstance. Stability is then defined as the extent to which the causal relationship $X \rightarrow Y$ holds in a variety of background circumstances. (One can think of background circumstances as circumstances not included in X and Y.) 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. Woodward argues convincingly that stability considerations play an important role in scientific practice, especially in selecting appropriate levels of causal representation and explanation.

There is also some indirect evidence for the role of stability in people's intuitive causal and explanatory judgments. Lombrozo (2010) found that people are more willing to consider relationships causal when an association involves a direct physical connection rather than double prevention, and – when double prevention is involved – are more inclined to regard an agent as a cause of an outcome when the action was intentional (vs. accidental). She argues that both effects could be due to a difference in the stability of the relevant relationships. Likewise, 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, 2016). A possible explanation suggested by Murray and Lombrozo 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. However, so far few direct investigations of the role of stability in lay causal judgments have been attempted (but see Gerstenberg et al., 2012), and none that appropriately control for such relevant features as number of intermediate causes and causal strength. In particular, to show that stability has an effect over and above causal strength, it is essential to consider cases where stability varies while causal strength (e.g., ΔP) is held fixed.

We conducted two experiments to investigate whether people are sensitive to stability considerations. Participants were presented with evidence suggesting either that a causal relationship holds in only one out of two kinds of circumstances, or that it holds in both kinds of circumstances. The causal strength of the relationship (for the full set of cases) was held fixed across the two conditions. In philosophy, the notion of stability has been applied to causal relations both between *types* (Woodward, 2010) and between *token* events (Woodward, 2006), and is held to be important both for *causal* and *explanatory* judgments (Woodward, 2010). To test for these different aspects and roles of stability, we asked participants to rate either causal or explanatory judgments at either the type or token level. (This also allowed us to ensure that any observed effect wasn't merely due to idiosyncrasies in the formulation of particular questions.) 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.

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 *D* (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 *D* is present and when it is absent. In the *moderated* condition, by contrast, the evidence suggested that *C* causes *E only* when *D* 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 its stability varied. The relationship was stable with respect to the *moderator variable* (presence or absence of *D*) in the non-moderated condition, but unstable with respect to this variable in the moderated condition.

Participants were asked to rate statements about the relationship between *C* and *E* in the overall population. As the causal strength of the relation was the same in both conditions, an effect of stability on causal and explanatory judgments should manifest itself in higher ratings in the stable (*non-moderated*) than in the unstable (*moderated*) condition.

Method

Participants One-hundred-eighty-two 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 memory check.

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 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. Participants were told that to test the hypothesis, the scientist performed an experiment, 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 x 2 covariation table cross-classifying zelmos based on whether they ate yonas or not, and whether they developed sore antennas or not. The numbers in the table were selected to support a causal strength with a ΔP of about .4 (range .39-.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

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)

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

	Causal judgment	Explanation judgment
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.*
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.*

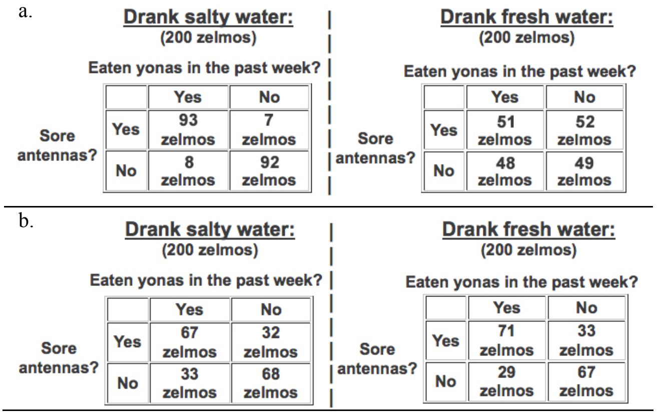
variable were always said to occur normally on the planet (e.g., in the wild, zelmos drink either fresh or salty water, depending on what's available). Luckily for the scientist, the moderator and cause variables varied orthogonally. Participants were told that "to see whether drinking salty water made 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 Figure 1).

We varied whether the split tables indicated a relationship that was *moderated* or *not moderated*. In the moderated cases (illustrated in Figure 1a), in one subgroup (salty water) the relationship between eating yonas and sore antennas was very strong (ΔP=.82-.85), while in the other subgroup (fresh water), the relationship nearly disappeared (ΔP= -.06-.01). In the non-moderated cases (Figure 1b), each of the split tables corresponded to relationships with a ΔP comparable to the ~.40 from the original, unsplit table. Importantly, the average strength of the relationship across the two split tables was the same in the moderated and non-moderated conditions, and equaled the strength of relationship in the first table that participants saw for each item (within .02 ΔP units).[1] 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 general, i.e., they stated a relationship between eating yonas and sore antennas without mentioning the kind of water the zelmo(s) in question drank.[2] 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) x 2 judgment (causal vs. explanatory) x 2 target (type vs. token) mixed design, with moderator manipulated within-subjects. The dependent variable was agreement with causal or explanatory claims, measured on a 1 (strongly disagree) to 7 (strongly agree) scale.

Results and Discussion

Our main question was whether relationships with known moderators support general causal and explanatory claims to the same extent as relationships without known moderators. A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) mixed ANOVA on ratings revealed a main effect of moderator: as shown in Figure 2, participants were significantly less likely to agree with claims about causal and explanatory relationships when a relationship was moderated than non-moderated, $F(1,178)=163.22$, $p<.001$, $\eta_p^2=.478$, 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 or interactions (all p's$\geq$ .211), suggesting that the effect of moderator was not itself moderated by the nature of the judgment (causal or explanatory, type or token).



a.

Drank salty water: (200 zelmos)

Eaten yonas in the past week?

		Yes	No
Sore antennas?	Yes	93 zelmos	7 zelmos
	No	8 zelmos	92 zelmos

Drank fresh water: (200 zelmos)

Eaten yonas in the past week?

		Yes	No
Sore antennas?	Yes	51 zelmos	52 zelmos
	No	48 zelmos	49 zelmos

b.

Drank salty water: (200 zelmos)

Eaten yonas in the past week?

		Yes	No
Sore antennas?	Yes	67 zelmos	32 zelmos
	No	33 zelmos	68 zelmos

Drank fresh water: (200 zelmos)

Eaten yonas in the past week?

		Yes	No
Sore antennas?	Yes	71 zelmos	33 zelmos
	No	29 zelmos	67 zelmos

Figure 1: Sample covariation matrices from Experiment 1: (a) split tables in the moderated condition, ΔP's=.85 and .00 (*M*=.42); (b) split tables in the non-moderated condition, ΔP's=.35 and .38 (*M*=.37).

[1] For each item, average ΔP's in the moderated and non-moderated conditions could differ slightly (by no more than .05 ΔP units). Importantly, the non-moderated condition strength never exceeded the moderated condition strength, which worked against our hypothesis. (This also holds for other metrics of causal strength computed over covariation tables, e.g. causal power, Cheng, 1997.)

[2] In both Experiments 1 and 2, an additional group of participants evaluated qualified causal and explanatory claims that specified the subgroup defined by the moderator variable, e.g., *For zelmos who drank salty water, eating yonas causes their antennas to become sore*. Participants in both experiments also evaluated counterfactual claims. Due to space limitations, here we focus on unqualified claims only, and we omit counterfactual ratings.

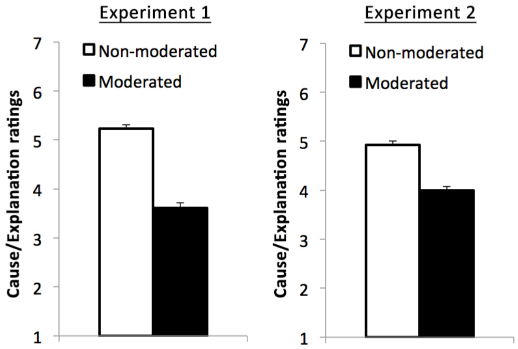


Figure 2: The effect of moderator on ratings of causal and explanatory relationships in Experiments 1 (left) and 2 (right). Error bars correspond to 1 SEM.

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

Experiment 2

The results of Experiment 1 provide initial evidence for an effect of stability (i.e., invariance across a range of circumstances) on causal and explanatory ratings. Yet these results are also amenable to an alternative interpretation. In Experiment 1, the moderated relationship had two characteristics. First, it was relatively unstable, in that it held in only one kind of circumstance described in the fictional world. The non-moderated relationship, by contrast, held in both kinds of circumstances. Second, the moderated relationship had a narrower *actual scope*, i.e., 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. By contrast, the non-moderated relationship held in the entire actual population. But stability and actual scope are distinct: an unstable relationship can have wide actual scope if the circumstance in which it holds happens to be frequent. This suggests an alternative explanation for the results of Experiment 1: moderated relationships could be penalized merely for their narrow actual scope. This alternative explanation also suggests that the penalty for moderated relationships could be a superficial pragmatic phenomenon: a generalization about a population could be infelicitous when it applies only to a small actual portion of the population. For instance, "having sex can cause deep vein thrombosis" is potentially *misleading* if the generalization only applies to women who become pregnant.

To address the possibility that our results are driven by a concern for actual scope rather than stability, we conducted a further experiment where in addition to varying the number of circumstances in which a causal relationship holds (thus investigating effects of stability), we 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. 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. By using such a formulation, which taps more directly into participants' beliefs about causal relationships as opposed to communication and language use, we hoped to address the possibility that the results of Experiment 1 were due to some pragmatic infelicity of our general (unqualified) claims.

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.

Materials, Design, and Procedure The materials, design and procedure were the same as in Experiment 1, with the following exceptions. First, 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 the research assistants' mistakes), 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 for which the causal relationship held) in the natural population and in the sample. This circumstance occurred in either 30% (*low frequency*), 50% (*medium frequency*), or 70% of cases (*high frequency*). Participants were told that the sizes of the groups were intentionally matched to the frequency of the enabling circumstance in the natural population.

To keep the mean strength of causal relationships (averaged across split tables) the same (ΔP=.31) in the moderated and non-moderated condition despite variation in the base rate

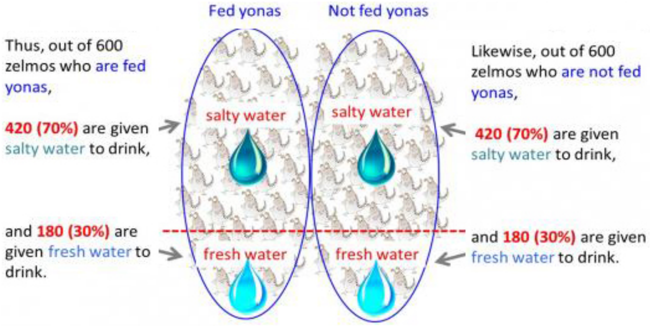

Fed yonas
Not fed yonas
Thus, out of 600 zelmos who are fed yonas,
420 (70%) are given salty water to drink,
and 180 (30%) are given fresh water to drink.
salty water
salty water
fresh water
fresh water
Likewise, out of 600 zelmos who are not fed yonas,
420 (70%) are given salty water to drink,
and 180 (30%) are given fresh water to drink.

Figure 4: Sample diagram provided to participants to illustrate the design of a hypothetical study (high-frequency moderator condition).

approximately equal to zero in one subgroup, the strength of the relationship in the subgroup with a causal association inevitably had to vary (ΔP=.97, ΔP=.61, and ΔP=.44 for low, medium, and high, respectively).

Thus, Experiment 2 had a 2 moderator (moderated vs. non-moderated relationship) x 2 judgment (causal vs. explanatory) x 2 target (type vs. token) x 3 moderator frequency (low 30%, medium 50%, high 70%) mixed design, with moderator manipulated within-subjects. The main dependent variables were agreement with causal or explanatory claims, measured on a 1 (strongly disagree) to 7 (strongly agree) scale.

Participants also answered 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.* Here we report the findings for judgments concerning the candidate cause and effect (e.g., eating yonas → sore antennas). To prevent participant fatigue given additional ratings, the number of items was reduced to two (see Table 1).

Results and Discussion

Main ratings. A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) x 3 moderator frequency (low, medium, high) mixed ANOVA on main ratings revealed two main effects. As predicted, moderated relationships were rated lower than non-moderated relationships $F(1,381)=79.19$, $p<.001$, $\eta_p^2=.172$, replicating the moderator effect from Experiment 1 (see Figure 2b). In addition, type ratings (M=4.63) were higher than token ratings (M=4.26, $F(1,381)=8.97$, $p=.003$, $\eta_p^2=.023$. There were no other main effects or interactions (all p's≥.154). Most notably, there was no effect of moderator frequency, $F(2,381)=1.60$, $p=.203$.

Ratings of causal structure and strength A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) x 3 moderator frequency (low, medium, high) mixed ANOVA on causal structure ratings revealed a significant main effect of moderator, $F(1,381)=56.49$, $p<.001$, $\eta_p^2=.129$. As shown in Figure 5, ratings were on average higher for the non-moderated relationship than for the moderated relationship. There was also a significant interaction between frequency and target, $F(2,381)=3.48$, $p=.032$, $\eta_p^2=.028$; there was a trend such that token ratings were lower than type ratings in the high frequency condition but not in others; however none of the simple effects were significant (p's≥.275).

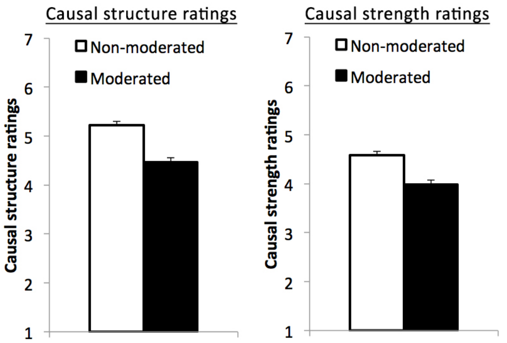


Figure 5: The effect of moderator on ratings of causal structure and strength in Experiment 2. Error bars correspond to 1 SEM.

An equivalent analysis of strength judgments also revealed a significant main effect of moderator, $F(1,380)=38.76$, $p<.001$, $\eta_p^2=.093$, with higher ratings for unmoderated than moderated relationships (Figure 5). No other effects reached significance.

In sum, Experiment 2 shows that stability affects causal judgments even controlling for causal strength, and that this influence is unlikely to reflect the frequency of the moderating circumstance or pragmatic considerations.

General Discussion

In two experiments, we document an important factor shaping people's assessments of causal 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 so far been largely unacknowledged in psychology (but see Gerstenberg et al., 2012, 2015). Experiments 1 and 2 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 2 also indicate that the effect of stability is not reducible to actual scope (that is, unstable causal relationships are not penalized merely because they hold in a smaller actual proportion of the population).

These findings point to an important limitation of the metrics of causal influence that have dominated the psychological research on causal reasoning and induction, such as ΔP or power PC. These measures track one aspect of causal relationships (their average strength in a population), but do not capture another important aspect that matters for causal assessment, namely the extent to which the relationship holds in a range of plausibly occurring background circumstances.

Our findings are related to the work of Liljeholm and Cheng (2007), who show that people can infer the presence of background factors interacting with a causal relationship based on differences in covariation across contexts. These findings are important in demonstrating that people are able to track the kind of evidence relevant to assessments of stability, which our experiments show will in turn affect the endorsement of general causal claims as well as ratings for causal structure and strength.

Our results also provide support for the *exportability* theory of explanation (Lombrozo & Carey, 2006) and causal ascriptions (Lombrozo, 2010). According to this theory, 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 provides more reliable opportunities for future prediction and intervention.

Our findings also suggest directions for future research. First, 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.

Second, what are the boundaries of the observed effects of stability? In a very different paradigm involving collisions among physical objects, Gerstenberg et al. (2012) found that the robustness of an outcome (whether a ball clearly or barely went through a gate) did not affect "cause" versus "prevent" judgments. However, it 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).

Third, is it possible (and necessary) to draw a line between the stability of a causal relationship with respect to *background circumstances* (defined loosely as circumstances not included in *cause* and *effect*) versus the stability of an outcome with respect to the *manner* in which the target cause occurs (Lombrozo, 2010), and/or to the *status of intermediate causes* in a causal chain (e.g., as in Gerstenberg et al., 2012, 2015)?

Fourth, can stability account for intransitivity in causal chains? For instance, it's reasonable to say that sex causes pregnancy, which 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 subsets of background circumstances for which the component relationships hold.

Fifth, how does the stability of a relationship across a range of circumstances relate to the degree of *guidance* it provides? Consider again the causal relationship between eating yonas and getting sore antennas in the case where it holds only in one background circumstance. One way to alleviate this instability is to explicitly build this 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 that eating yonas causes sore antennas – 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 is therefore exportable in the sense that it contains conditions for application, whether or not those conditions hold widely. Thus one question is whether people are sensitive to considerations of *guidance* when assessing unstable relationships, and how guidance (achieved by building in background circumstances) differs from offering an explanation or causal claim with the enabling conditions instead identified as additional, interacting causes. The roles of stability and guidance in causal ascription and explanation are ripe for further investigation.

Acknowledgments

This work was supported by the Varieties of Understanding Project, funded by the John Templeton Foundation.

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.

Cheng, P. (1997). From covariation to causation: A theory of causal power. *Psychological Review, 104,* 367-405.

Gerstenberg, T., Goodman, N., Lagnado, D.A., & Tenenbaum, J.B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In Miyake, N., Peebles, D., and Cooper, R. P., (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 Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock,T., and Maglio, P. P., (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society, pp.782–787, Austin, TX. Cognitive Science Society.

Griffiths, T.L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

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.

Liljeholm, M., & Cheng, P. (2007). Coherent generalization across contexts. *Psychological Science, 18,* 1014-1021.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions,,functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303-332.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167-204.

Murray, D., & Lombrozo, T. (2016). Effects of manipulation on attribution of causation, free will, and moral responsibility. *Cognitive Science*,.doi: 10.1111/cogs.12338.

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science, 39(6),* 1320-1347.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review, 115*, 1-50.

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

Woodward, J. (2016). The problem of variable choice. *Synthese, 193*, 1047-1072.