**Online Supplement A**

In this supplement we further consider the relationship between our studies and the Power PC framework (Cheng, 1997).

In our set of studies we pursue a different project from that pursued by Cheng and colleagues. We focus on how people evaluate causal and explanatory generalizations, while research on causal power (and related work, such as Spellman's, 1996a, 1996b) focuses on the problem of causal learning: how people infer a causal relationship from covariation data by isolating the influence of one variable from potential confounds. The latter project has a more pronounced normative element: it states what focal sets people *should* look at in order to evaluate a potential generative or preventive causal relationship, and which computations are appropriate to disentangle a target causal relationship from potential confounds.

By contrast, the notion of stability (at least as developed by Woodward, 2006, 2010) is intended to distinguish between various grades of (real) causal relationships – the hypothesis being that stable causal relationships are in some sense `better' or more useful than relationships of causal influence that hold only in very specific circumstances. Correspondingly, our project is not to examine how people infer the existence of a causal relationship from statistical data, but whether they make distinctions between causal relationships based on the extent to which they hold across various circumstances. In accordance with this objective, our aim is to make sense of the everyday practice of making unqualified causal claims, such as "aspirin reduces fever," or asking about unqualified causal relationships, such as “does stress reduce academic performance?” We regularly ask – and answer – such questions, even if we are aware of a variety of potentially relevant focal sets, and know or suspect that the relationship may be unstable across those sets (people with different health conditions / genotype / diet / age, etc.). We also often reason about token cases when values of potential moderating variables are unknown.[1] The pervasiveness of generic causal statements in daily communication – “aspirin reduces fever,” “ticks cause Lyme disease” - suggests that considering such causal generalizations is a natural psychological process.

While the causal power project provides a very sophisticated account of how people evaluate simple and/or conjunctive causal powers from data (Cheng, 1997, 2000; Novick & Cheng, 2004), it does not speak to how people evaluate summary statements about cause C that generalize across its simple and conjunctive powers. In other words, the causal power framework does not address how people would deal with causal generalizations (e.g., “eating yonas causes sore antennas”), which do not reduce to either simple or conjunctive causal powers, and could potentially draw upon elements of both. Likewise, it does not offer a formalization of stability, which tracks the extent to which the target cause interacts with background variables (which is different from the causal power of a conjunctive cause containing the target cause; rather, in the language of an ANOVA, stability would track how “qualified” a main effect is).

[1] The fact that we find effects of stability even for token-level causal and explanatory judgments suggests that people may consult population-level stability when evaluating specific cases. See Woodward (2006) on how the notion of stability applies to causal relations between token events.

Of course, one could argue that people in fact do not consider such summary causal generalizations, and that in our study they conceptualize the moderator variable in one of the ways already formalized within the Power PC approach. We consider three such ways and discuss whether it is possible to account for our findings using the theoretical machinery from the Power PC framework / Probabilistic Contrast Model (Cheng, 1993, 1997; Cheng & Novick, 1990; Novick & Cheng, 2004).

### 1. Participants see the moderator variable as an alternative generative cause, and rate the simple causal power of the target cause in the absence of the moderator.

Based on the data provided to participants, the moderator variable does not meet the criteria for an alternative cause under Cheng and Novick's (1991, 1992) definition, which stipulates that it should covary with the effect in a focal set other than the one currently considered by the reasoner[2] (in fact, the available evidence suggests that the moderator variable *does not* covary with the effect in the total currently available dataset). But let's assume that participants nevertheless treat the moderator as an alternative cause. According to the Power PC account, to evaluate the causal power of a generative cause people prefer to consider the focal set(s) where alternative causes are absent (Cheng & Holyoak, 1995). In other words, if people consider salty water as an alternative cause of sore antennas, when asked to evaluate a claim about yonas causing sore

[2] For the same reason, the moderator variable does not qualify as an enabler under Cheng, Park, Yarlas, & Holyoak's (1996) definition: "…candidate *i* [is] an enabling condition for a cause *j* if *i* is constantly present in a reasoner's current focal set but covaries with the effect in another focal set, and *j* no longer covaries with the effect in a focal set in which *i* is constantly absent" (p.313).

antennas they should base their judgments on the focal set where salty water is absent – that is, they should respond as if they were asked to rate the causal relationship between yonas and sore antennas conditional on drinking fresh water. To evaluate this possibility, we can compare participants' main type-level ratings (e.g., of the causal generalization "For zelmos, eating yonas causes their antennas to become sore") to their ratings of the causal/explanatory claims explicitly conditionalizing the relationship on the "absence" of the moderator variable[3] (what we call a "low-moderator" group, e.g., "For zelmos *who have been drinking fresh water*, eating yonas causes their antennas to become sore"). This comparison for ratings of the moderated relationships from Experiment 1 indicates that conditional (qualified) ratings for the "low-moderator" group (*M=2.65)* were significantly lower than the main set of unqualified ratings in the moderated condition (*M*=3.62, *t*(180)=4.61, *p*<.001, Cohen's *d*=.69). The same difference held for qualified and unqualified structure and strength ratings in Experiment 2: both causal structure and causal strength ratings of moderated relationships were lower for statements about the "low-moderator" subgroup ($M_{\text{structure}}$=3.15, $M_{\text{strength}}$=2.78) than for unconditional statements ($M_{\text{structure}}$=4.53, $M_{\text{strength}}$=3.89; $t_{\text{structure}}$(196)=9.65, *p*<.001, Cohen's *d*=.67, $t_{\text{strength}}$(196)=7.65, *p*<.001, Cohen's *d*=.55; see online Supplement C for sample ratings).[4]

[3] Our items varied in how straightforward it was to determine which of the conditionalized focal sets (covariation tables split by the moderator variable) represented the relationship *in the absence* of the moderator variable: from clear cases (smoke exposure occurred vs. did not) to less straightforward cases (hot vs. cold temperature; salty vs. fresh water – depending on the ambient temperature and prevalent chemical composition of water, either value of could be seen as setting the moderator variable to ON). The fact that we observed little inter-item variability in results speaks against the claim that people were looking exclusively for the focal sets *in the absence* of alternative factors.

[4] A comparison of unqualified causal and explanatory claims in Experiment 2 with qualified causal and explanatory claims collected in an additional experiment reported in the Online Supplement F is of limited informativeness, due to the differences in the way the causal strength was controlled across frequency conditions in these experiments (see Online Supplement F). A comparison of qualified and unqualified causal

This suggests that participants were not treating our causal generalizations the way they would evaluate the causal powers of generative causes on the Power PC account.

Furthermore, if participants interpreted all of our causal generalization evaluation tasks as requests to report back simple causal powers (in the absence of alternative causes), then when asked to evaluate claims about the moderating variable (e.g., drinking salty water), they should have concluded that in the moderated condition, the moderator variable (drinking salty water) *suppresses* the outcome (sore antennas), whereas in the non-moderated condition, it has no effect on the outcome. To see why this is the case, consider the data participants observed in the moderate frequency condition (Figure S1, panels a and b); pulling the data from the right columns of each table (shaded), we obtain covariation tables for drinking salty water and developing sore antennas in the absence of eating yonas (Figure S2, panels a and b).

If participants were evaluating the simple causal power of the moderator variable to produce the outcome in the absence of alternative causes, they should have said that the relationship between the moderator variable and the outcome is absent in the non-moderated condition, and is negative in the moderated condition (the same pattern would hold for the low and high frequency conditions). In contrast to this prediction, as reported in footnote 10, in both the moderated and non-moderated conditions the structure and strength ratings of the moderator variable → effect relationships were significantly above the lowest scale endpoint corresponding to the absence of a relationship / a very weak relationship. Furthermore, participants gave higher structure and strength ratings to the

and explanatory claim ratings from Experiment 3 is not informative due to the way the data were presented in the form of summary statements. Nevertheless, the predicted difference held even for all these judgments.

moderator variable → effect relationship in the moderated than non-moderated condition, which deviates even further from the prediction above.

a. Non-moderated condition

Drank salty water
(600 zelmos)

| | | Eaten yonas in the past week? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 190 | 98 |
| | No | 110 | 202 |

Drank fresh water
(600 zelmos)

| | | Eaten yonas in the past week? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 198 | 105 |
| | No | 102 | 195 |

b. Moderated condition

Drank salty water
(600 zelmos)

| | | Eaten yonas in the past week? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 241 | 60 |
| | No | 59 | 240 |

Drank fresh water
(600 zelmos)

| | | Eaten yonas in the past week? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 151 | 151 |
| | No | 149 | 149 |

*Figure S1*. Sample covariation data between the target cause and outcome, split by the moderator variable, as presented to participants in Experiment 2, medium frequency condition, non-moderated (a) and moderated (b) conditions. Column shading added for illustrative purposes.

a. Non-moderated condition

Did not eat yonas

| | | Drank salty water? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 98 | 105 |
| | No | 202 | 195 |

b. Moderated condition

Did not eat yonas

| | | Drank salty water? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 60 | 151 |
| | No | 240 | 149 |

*Figure S2.* Subset of data from Figure S1 rearranged to represent the relationship between the moderator variable and the outcome in the absence of the alternative (target) cause in the non-moderated (a) and moderated (b) conditions.

### 2. Participants combine the target cause and the moderator variable into an ***interactive cause*.**

The Causal power approach offers nice machinery to describe how one would calculate the causal power of a conjunctive cause (e.g., treating “eating yonas and drinking salty water” as a composite cause which is present only when both elements are present, and absent otherwise), and it offers a clear account of how one would calculate a simple effect corresponding to each component of an interactive cause (e.g., “eating yonas,” “drinking salty water”). Could we account for our results by assuming that to evaluate our causal and explanatory claims about the target cause (“eating yonas”), people report back the causal power of the interactive cause, or some weighted combination of the interactive and simple causal powers? To evaluate this possibility, we calculated the mean[5] simple causal power of the target cause variable (as formalized in Cheng, 1997) and the interactive power of the conjunctive cause (Novick & Cheng, 2004) consisting of the target cause and the moderator variables from the co-variation tables presented to participants in the moderated and non-moderated condition, separately for each moderator frequency setting in Experiment 2. The resulting simple and conjunctive causal power profiles are shown in Figure S3, panels (a) and (b). Panel (c) of Figure S3 shows the pattern of ratings we would expect to observe if participants simply computed an unweighted mean of the simple and conjunctive powers. For comparison, panel (d) of Figure S3 reproduces the actual causal and explanation ratings participants gave in Experiment 2 (the rest of the ratings are shown in Figure 4 in the main manuscript). It is clear that neither simple nor conjunctive causal powers, nor some weighted or unweighted combination of the two (with a single set of weights systematically applied *across all experimental conditions*), could produce the pattern of results we see in our data, in particular reproducing the lack of interaction between the moderator and frequency factors observed in our data.

[5] We averaged across the two items used in this Experiment, zelmos and drols. Restricting this analysis to either item produces identical results.

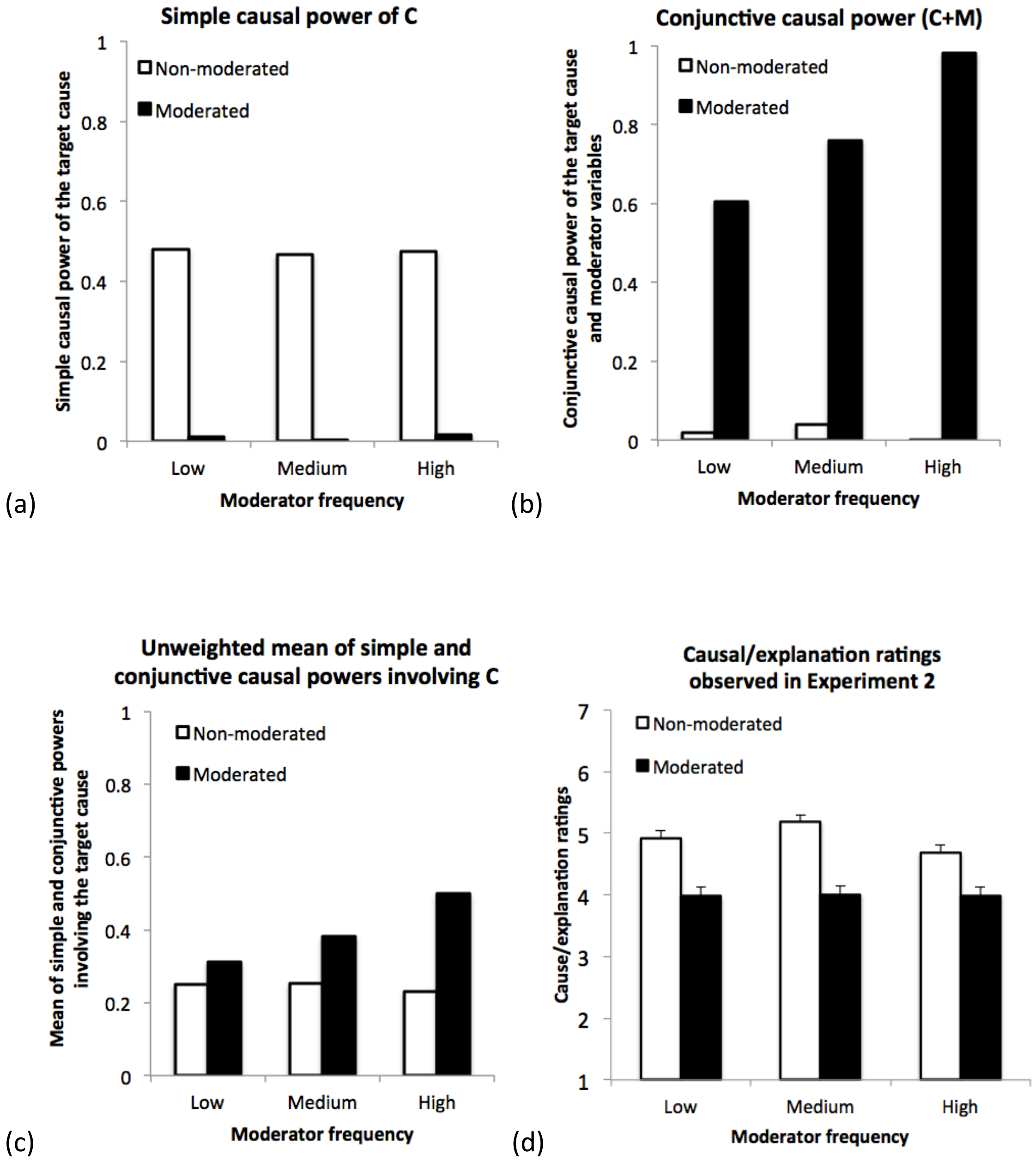


*Figure S3*. Mean simple causal power of the target cause C, e.g. "eating yonas" (a) and causal power of the conjunction of the target cause C and moderator variable M (b) as a function of moderator and frequency condition, calculated from the covariation data provided to participants in Experiment 2. Panel (c) shows the unweighted average of the simple and conjunctive powers across the experimental conditions. For comparison, panel (d) reproduces the pattern of results observed in Experiment 2.

**3. Participants see the moderator variable as merely a way to partition data into multiple focal sets.**

Cheng and colleagues acknowledge the possibility that people may rely on multiple focal sets in estimating causal power (either when multiple informative focal sets are available, or when all available focal sets are flawed), but do not say much about the resulting behavior. For example, Cheng, Park, Yarlas, and Holyoak (1996) state that “If the causal powers revealed in multiple informative sets conflict, […] reasoners would have to either withhold judgment or resolve the conflict *in some way*” (p. 324, our emphasis; see also Cheng, 1997, p. 377). Although our proposal does not offer a quantitative model of the way in which such conflicts will be resolved, it says something more specific about how such cases will be evaluated: the causal generalization across conflicting sets will be seen as less appropriate than a causal generalization across non-conflicting sets, even if the average causal strength does not differ between the two scenarios.

We believe it’s less likely that people would treat our scenarios as describing flawed focal sets, as they would when there are confounds or ceiling effects. But if they did, we could attempt to account for our results based on Cheng and Holyoak’s (1995) “mixture-of-focal-sets” hypothesis, which states that when some evidence relevant to assessing conditional dependencies is missing, causal ratings may be based on multiple contingencies, including mixtures of conditional and unconditional contingencies. Cheng and Holyoak make no claims about the quantitative mapping between multiple contingencies and subjects’ responses other than “subjects’ causal estimates will increase monotonically with a nonnegatively weighted function of the contingency values of their focal sets. Individual subjects may compute and integrate multiple contingencies for a cue (e.g., by simple averaging). Alternatively, each subjects may use only one focal set, but different subjects may use different focal sets, in which case the mean ratings may mask distributions that are in fact multimodal" (p. 286). As we show in Supplement C, neither weighted nor unweighted combinations of the two focal sets (split by the moderator variable; e.g., showing the relationship between eating yonas and sore antennas given salty vs. fresh water) can account for our pattern of results in Experiment 2. The overall dataset, unsplit by the moderator variable, had the causal strength of the target cause (e.g. yonas) equated across the moderated and non-moderated conditions, so averaging one of the focal sets with a constant could not produce the observed difference either.

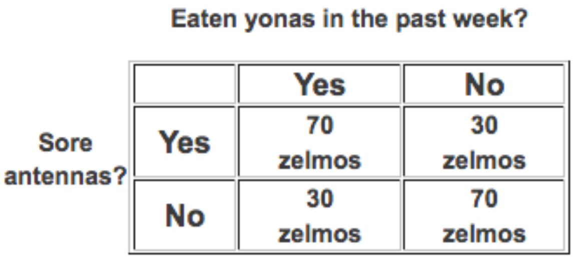
In sum, the only plausible focal sets participants could have relied on are the two sets conditionalized on the moderator, plus the overall set. It appears that calculating causal power across a weighted combination of these focal sets (including the extreme weighting schemes ignoring one of the sets) can not account for our results.

**Online Supplement B**

Experiment 1 sample trial (moderated condition, type-level causal judgments)

You are a scientist on planet Zorg. You study a lizard-like species called the zelmo, which can drink either fresh water or salt water, depending on what happens to be available.

Zelmos sometimes have sore antennas. You are investigating the hypothesis that eating a kind of plant called a yona may be causally related to the development of sore antennas in zelmos. To test the hypothesis, you perform an experiment. First, you select a random sample of 200 zelmos. Then you randomly assign these 200 zelmos to one of two diet plans. For a full week the 100 zelmos in the first group eat a diet that includes yonas every day. The 100 zelmos in the second group eat the same kinds of foods as the first group for a week, but without any yonas. At the end of the week, you check how many of the zelmos in each group have developed sore antennas. Here are the results:

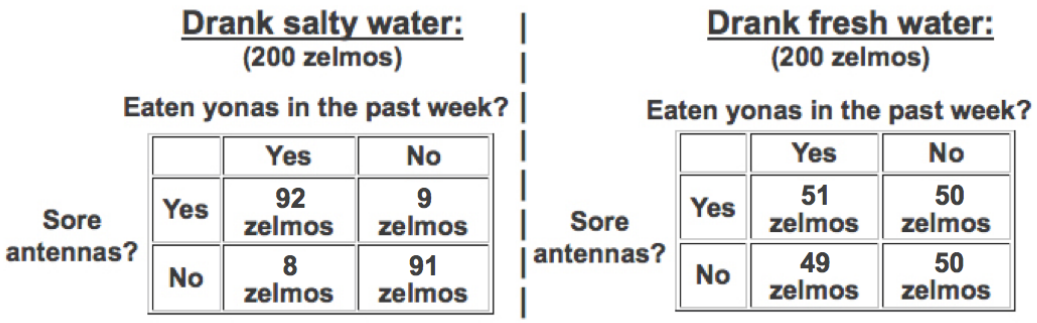


Eaten yonas in the past week?

|  |  | Yes | No |
|---|---|---|---|
| Sore antennas? | Yes | 70 zelmos | 30 zelmos |
|  | No | 30 zelmos | 70 zelmos |

As a follow-up, you decide to conduct a second experiment with an additional 400 zelmos. As in your first experiment, 200 are randomly assigned to eat a diet containing yonas for a week, and 200 are assigned to eat a diet without yonas.

After the experiment, you discover that there was a miscommunication between your research assistants. Half believed the zelmos should be given salty water, and half thought the zelmos should only receive fresh water. So half of the 200 zelmos who ate yonas had salt water and half did not, and half of the zelmos who did not eat yonas had salt water and half did not. To see whether drinking salt 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. Here are the results:



Drank salty water:
(200 zelmos)

Eaten yonas in the past week?

|  |  | Yes | No |
|---|---|---|---|
| Sore antennas? | Yes | 92 zelmos | 9 zelmos |
|  | No | 8 zelmos | 91 zelmos |

Drank fresh water:
(200 zelmos)

Eaten yonas in the past week?

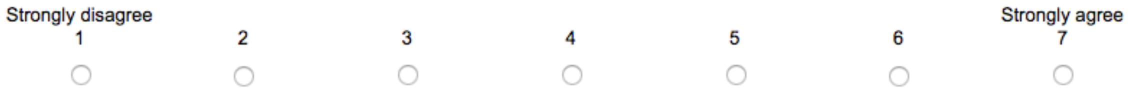
|  |  | Yes | No |
|---|---|---|---|
| Sore antennas? | Yes | 51 zelmos | 50 zelmos |
|  | No | 49 zelmos | 50 zelmos |

The tables reveal that the data pattern looks very *different* 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 the patterns are.

--------------------------------------------------

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.



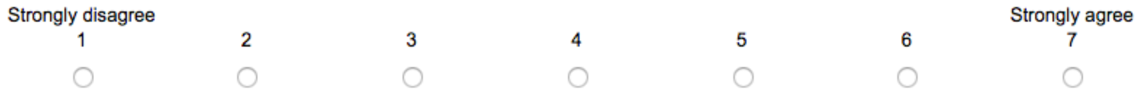
| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly disagree | | | | | | Strongly agree |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

[new screen]

**Here is a sample of 100 zelmos with sore antennas from your second experiment. These zelmos were fed yonas and drank salty water during the experiment.**

**Please tell us how much you agree with the following statement:**

**Had these zelmos eaten yonas but not drunk salty water, their antennas would still have become sore.**



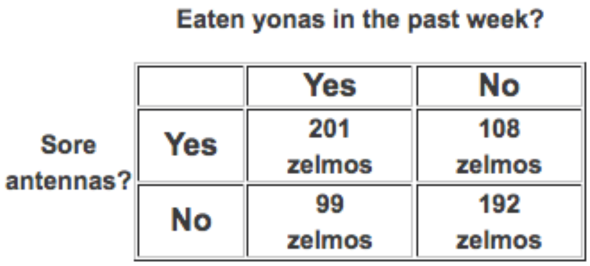
| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly disagree | | | | | | Strongly agree |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**Online Supplement C**

Experiment 2 sample trial (moderated relationship; high frequency condition; main rating: type-level explanatory judgment)

You are a scientist on planet Zorg. You study a lizard-like species called the zelmo, which drinks either fresh water or salty water, depending on what is available in its environment.

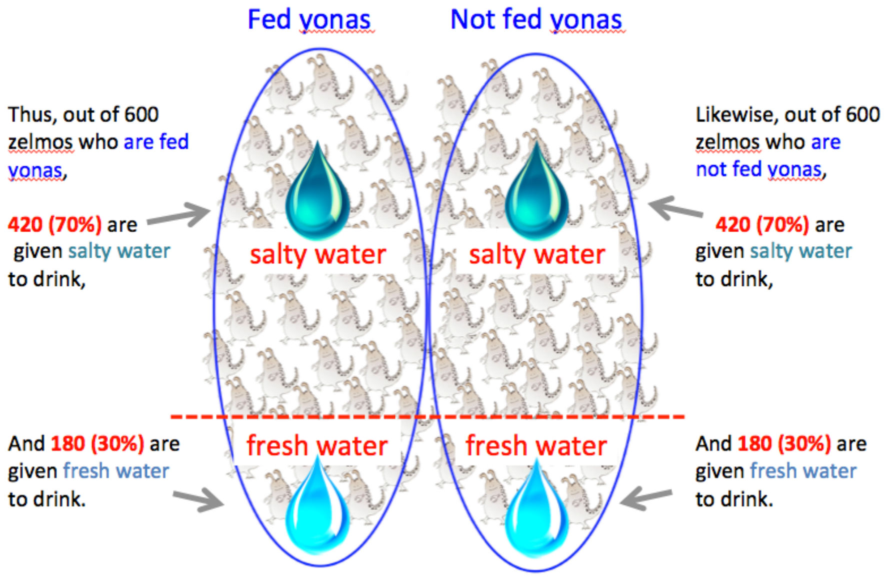
Zelmos sometimes have sore antennas. You are investigating the hypothesis that eating a kind of plant called a yona may be causally related to the development of sore antennas in zelmos. To test the hypothesis, you perform an experiment. First, you select a random sample of 600 zelmos. Then you randomly assign these 600 zelmos to one of two diet plans. For a full week the 300 zelmos in the first group eat a diet that includes yonas every day. The 300 zelmos in the second group eat the same kinds of foods as the first group for a week, but without any yonas. At the end of the week, you check how many of the zelmos in each group have developed sore antennas. Here are the results:



| | | Eaten yonas in the past week? | |
|---|---|---|---|
| | | Yes | No |
| Sore antennas? | Yes | 201 zelmos | 108 zelmos |
| | No | 99 zelmos | 192 zelmos |

As a follow-up, you decide to conduct a second experiment with an additional 1200 zelmos. You randomly select half (600) of these zelmos and assign them to a diet containing yonas for a week. The other half are assigned to eat a diet without yonas. But this time, you *decide to check whether drinking salty water makes a difference* to the effects of yonas on sore antennas.

There are approximately 100,000 zelmos on Zorg. In the wild, about 30,000 of them (30%) drink fresh water, and 70,000 of them (70%) drink salty water. *You want to make sure that in your sample of zelmos, the natural proportion of drinking salty vs. fresh water is represented accurately.* So you randomly select 420 zelmos from the group eating yonas for a week (that's 70% of this group) and give them salty water to drink that week, and the other 180 zelmos from the group eating yonas (that's 30% of this group) are given fresh water to drink that week. You do the same thing for the 600 zelmos in the group eating a diet without yonas.



Thus, out of 600 zelmos who are fed yonas,

420 (70%) are given salty water to drink,

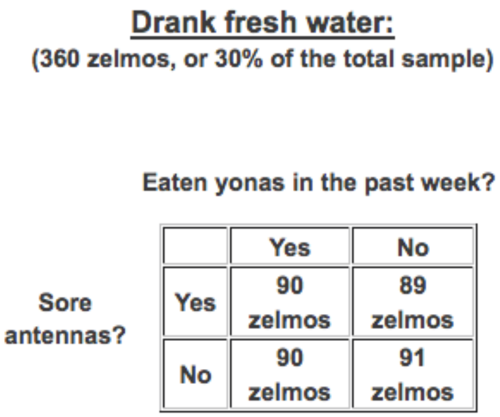
And 180 (30%) are given fresh water to drink.

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.

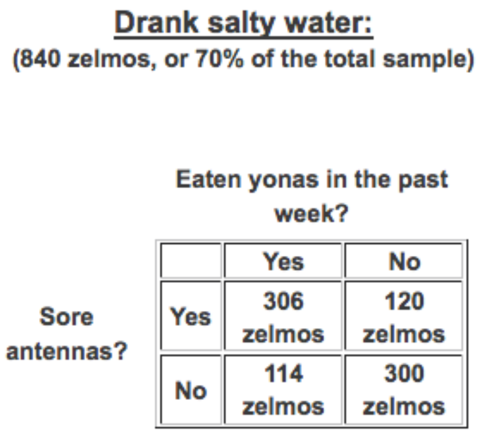
Here are the results of the experiment:



**Drank fresh water:**
(360 zelmos, or 30% of the total sample)

Eaten yonas in the past week?

| Sore antennas? | | Yes | No |
|---|---|---|---|
| | Yes | 90 zelmos | 89 zelmos |
| | No | 90 zelmos | 91 zelmos |



**Drank salty water:**
(840 zelmos, or 70% of the total sample)

Eaten yonas in the past week?

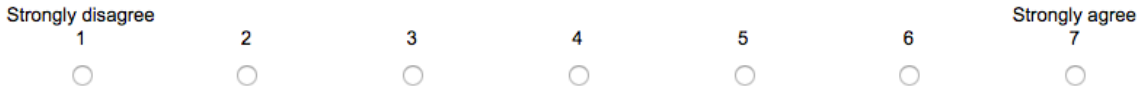
| Sore antennas? | | Yes | No |
|---|---|---|---|
| | Yes | 306 zelmos | 120 zelmos |
| | No | 114 zelmos | 300 zelmos |

The tables reveal that the data *pattern* looks very ***<u>different</u>*** 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 the patterns are.

- - - *[new screen]* - - -

How much do you agree with the following explanation of why zelmos' antennas become sore?:

For zelmos, antennas become sore because of eating yonas.



| Strongly disagree<br>1 | 2 | 3 | 4 | 5 | 6 | Strongly agree<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

- - - *[new screen; all the subsequent questions are presented one per page]* - - -

Next you will be asked some questions about **existence** and **strength** of causal relationships. Do you remember the difference between the two?

Reminder:

You will answer **existence**-questions on a scale from Not at all likely (1) to Very likely (7). For example, a causal relationship is likely to exist between wearing thin socks and getting a cold, but unlikely to exist between wearing thin socks and seeing an airplane.

You will answer **strength**-questions on a scale from Very weak relationship (1) to Very strong relationship (7). For example, the causal relationship between wearing thin socks and getting a cold is on the weaker side, but a causal relationship between putting your hand on a hot stove and getting burnt is on the stronger side.

**In your opinion, how likely is it that there is some *causal relationship* between eating yonas and having sore antennas?**

| Not at all likely<br>1 | 2 | 3 | 4 | 5 | 6 | Very likely<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

[the following question was presented only if the previous rating was 2 or higher]

**If there is a causal relationship between eating yonas and having sore antennas, how *strong* do you think it is?**

| Very weak relationship<br>1 | 2 | 3 | 4 | 5 | 6 | Very strong relationship<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**In your opinion, how likely is it that among zelmos who drink fresh water there is some *causal relationship* between eating yonas and having sore antennas?**

| Not at all likely<br>1 | 2 | 3 | 4 | 5 | 6 | Very likely<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

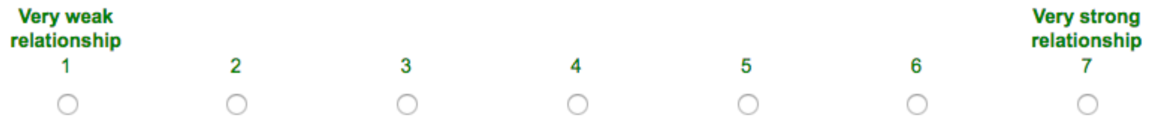
**If there is a causal relationship between eating yonas and having sore antennas among zelmos who drink fresh water, how *strong* do you think it is?**

| Very weak relationship<br>1 | 2 | 3 | 4 | 5 | 6 | Very strong relationship<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**In your opinion, how likely is it that among zelmos who drink salty water there is some *causal relationship* between eating yonas and having sore antennas?**

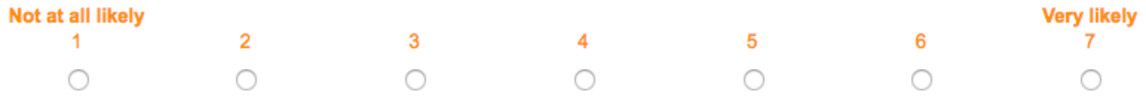
| Not at all likely<br>1 | 2 | 3 | 4 | 5 | 6 | Very likely<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**If there is a causal relationship between eating yonas and having sore antennas among zelmos who drink salty water, how *strong* do you think it is?**



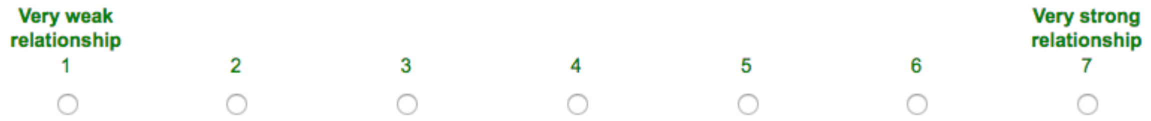
| Very weak relationship<br>1 | 2 | 3 | 4 | 5 | 6 | Very strong relationship<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**In your opinion, how likely is it that there is some *causal relationship* between drinking salty water and having sore antennas?**



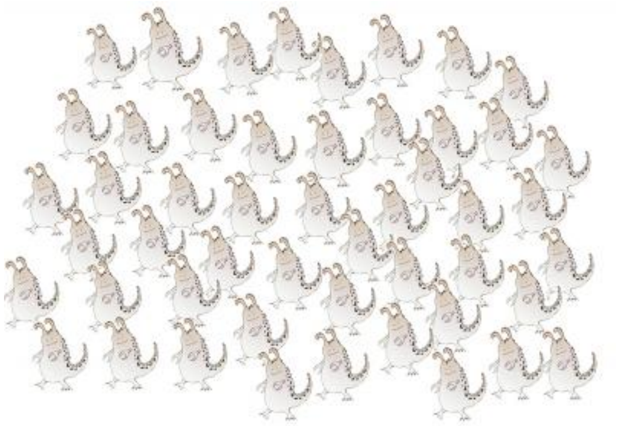
| Not at all likely<br>1 | 2 | 3 | 4 | 5 | 6 | Very likely<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**If there is a causal relationship between drinking salty water and having sore antennas, how *strong* do you think it is?**



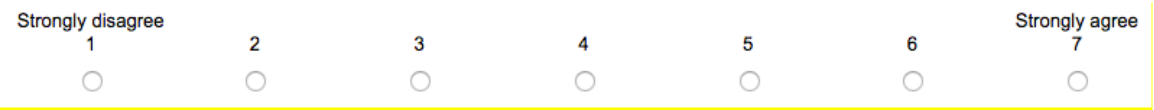
| Very weak relationship<br>1 | 2 | 3 | 4 | 5 | 6 | Very strong relationship<br>7 |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**Here is a sample of 50 zelmos with sore antennas from your second experiment. These zelmos were fed yonas and drank salty water during the experiment.**



**Please tell us how much you agree with the following statement:**

**Had these zelmos eaten yonas but not drunk salty water, their antennas would still have become sore.**



| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly disagree<br>1 | 2 | 3 | 4 | 5 | 6 | Strongly agree<br>7 |
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**Online supplement D**

In this Supplement we report an additional experiment evaluating an alternative explanation for differences across moderated and unmoderated conditions.

We attribute the observed differences between moderated and unmoderated conditions to stability, but an alternative explanation is that is that when participants receive information about a causal system that requires them to revise their causal model, as they must in our moderated conditions, they become less confident about causal beliefs that they previously held.[6] That is, the drop in ratings for the target cause (e.g., eating yonas) could result from the fact that when participants in the moderated condition see the split covariation tables, they learn that an additional factor (e.g., drinking salty water) is causally relevant to the outcome (e.g., sore antennas). This new information (that salty water is causally relevant; not the more specific claim that it moderates the relationship between yonas and sore antennas) might lead them to revise their belief about the causal role of eating yonas (the original cause) by reducing their confidence in the corresponding causal generalization. In contrast, in the non-moderated condition, the split covariation tables do not reveal any new information that might prompt participants to revise their beliefs about what causes the outcome. Thus, instead of tracking stability, participants could be responding to evidence that their causal knowledge is inaccurate or incomplete.

To address this possibility, we ran a modified version of Experiment 1 by adding a condition in which the second variable was presented as an independent cause of the outcome, rather than as a moderator. To illustrate this condition (“independent cause”),

[6] We are grateful to Jonas Nagel for suggesting this possibility.

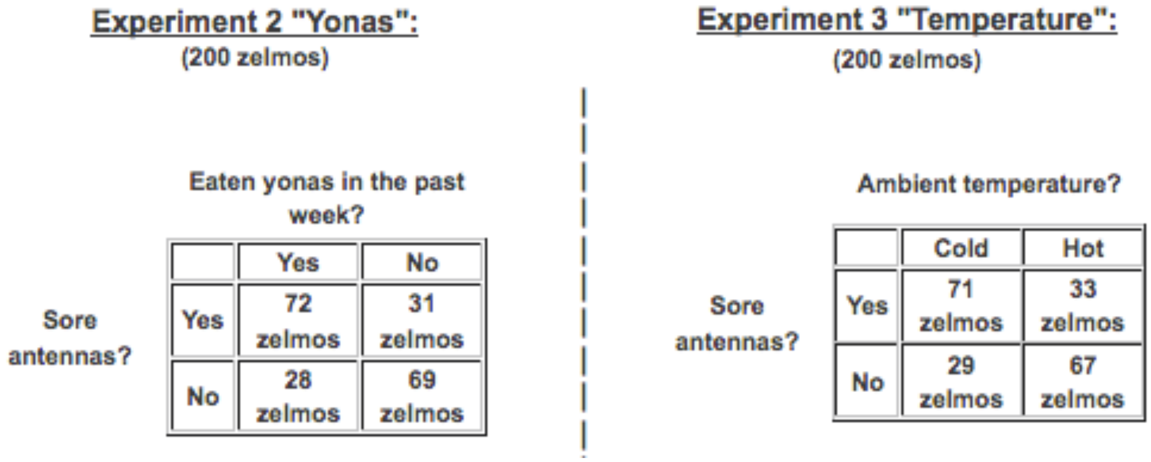
imagine a situation in which a person first learns that drunk driving causes accidents, and then learns that texting while driving independently causes accidents. Our original scenarios (“potential moderator”) instead correspond to a situation in which a person first learns that drunk driving causes accidents, and then that texting moderates this effect, but does not *independently* lead to accidents. In both cases, participants receive new information about the causal system, but only in the moderated condition is the original causal relationship unstable. If the drop in causal and explanatory ratings for the original cause is due to receiving new causal information (or specifically learning about alternative independent causes), we should observe the same pattern both in the “independent cause” and the “moderator” conditions. By contrast, if the effect is due to receiving new causal evidence that a relationship varies across background circumstances, in particular, then the “independent cause” condition should not mimic the “moderator” condition.

In our additional experiment, we thus varied whether the second variable was featured as a potential moderator or an independent cause, and whether it was causally active (i.e., moderated / independently caused the effect) or inert (i.e., did not moderate / did not independently cause the effect). Both manipulations were between subjects. The “potential moderator” condition was the same as in Experiment 1 with the following changes. First, to examine the generality of the effect, instead of “drinking salty water,” we used “ambient temperature” as the moderator variable: due to a miscommunication between research assistants in the vignette, half of the zelmos stayed in hot enclosures, and the other half stayed in cold enclosures. Second, for the sake of compatibility with the “potential independent cause” condition, the numbers in the split tables were adjusted slightly, resulting in ΔP's=.41 and .38 (*M*=.40) in the non-moderated condition, and ΔP's=.79 and .01 (*M*=.40) in the moderated condition. Third, at the end of the scenario we added a statement drawing participants' attention to the lack of independent causal influence between the moderator variable and the effect: "Overall, the rate at which zelmos kept in cold and hot areas got sore antennas did not appear to differ." Fourth, only the "zelmo" vignette was used.

In the "potential independent cause" condition, the first part of the vignette about hypothetical "Experiment 1," including the first covariation table, was identical to the "potential moderator" condition. But the rest of the vignette described two additional experiments: "Experiment 2" repeated "Experiment 1," but "Experiment 3" introduced a new variable: "In Experiment 3, you want to investigate a new relationship: between ambient temperature and sore antennas. You want to see if the temperature may be causally related to the development of sore antennas in zelmos. You select yet another sample of 200 zelmos. This time, 100 are randomly assigned to *stay in cold enclosures* for a week, and 100 are assigned to *stay in hot enclosures* for a week." It was made clear that Experiments 2 and 3 were run separately, and in Experiment 2 all the zelmos were kept at normal room temperature, while in Experiment 3 none of the zelmos were given any yonas. Participants then saw two co-variation tables side by side, each showing results from *one* experiment[7]. The novel variable co-varied with the outcome in the "causally

[7] Since the specific purpose of this experiment was to test the alternative explanation that new information about an additional independent cause of the same effect is sufficient to produce the effect we attribute to stability, and since we already examined the effect of pitting evidence of stability vs. evidence of instability in the main set of studies, the independent cause condition offered no evidence regarding the stability of the independent causes (i.e., it was not possible to evaluate stability from the provided covariation tables). Our test of the target hypothesis holds regardless of the assumptions participants might make about the stability

active" condition (ΔP= .38; see Figure S4), but not in the "causally inert" condition (ΔP= .01). The final statement drew attention to the causal status of the second variable: "The rates at which zelmos kept in hot and cold areas got sore antennas did not appear [appeared] to differ: [a higher number of zelmos developed sore antennas in cold enclosures than in hot enclosures]."



Experiment 2 "Yonas":
(200 zelmos)

Sore antennas? / Eaten yonas in the past week?

| | Yes | No |
|---|---|---|
| Yes | 72 zelmos | 31 zelmos |
| No | 28 zelmos | 69 zelmos |

Experiment 3 "Temperature":
(200 zelmos)

Sore antennas? / Ambient temperature?

| | Cold | Hot |
|---|---|---|
| Yes | 71 zelmos | 33 zelmos |
| No | 29 zelmos | 67 zelmos |

*Figure S4.* Sample data presented to participants in Experiment 1b, "potential alternative cause" condition, second variable status: causally active.

of the mentioned relationships. It is plausible to expect that in the absence of evidence for or against moderation, participants would not spontaneously postulate moderation. However, even if they do so, that would not produce the predicted interaction: if they indiscriminately assume moderation regardless of whether the additional independent cause is active or inert, that would produce an overall drop in ratings in both the active and inert independent cause conditions; if the propensity to postulate moderation varies depending on the active/inert status of the additional independent cause, it would either eliminate the predicted interaction while sustaining the main effect of status (if participants are more likely to postulate moderation for an active than inert independent cause), or produce an interaction pattern opposite to what we predicted (if participants are more likely to postulate moderation for an inert rather than active independent cause – a less inviting but not unreasonable move). Thus our study design is appropriate for ruling out the alternative explanation in terms of belief revision in light of new information about an independent cause.

Following the vignette, all participants rated their agreement with the causal statement about the original cause: “For zelmos, eating yonas causes their antennas to become sore” on a 1 (strongly disagree) to 7 (strongly agree) scale.

A 2 (second variable type: potential moderator, potential independent cause) x 2 (second variable status: causally active (i.e. moderates / independently causes the effect), causally inert (i.e. does not moderate / independently cause the effect)) ANOVA on causal ratings from 166 participants (additional 17 excluded after failing comprehension checks) revealed two main effects (see Figure S5). First, there was a main effect of second variable type, $F(1,162)=9.70$, $p=.002$, $\eta_p^2=.056$, with lower ratings when the second variable was a moderator. Second, there was a significant main effect of second variable status, $F(1,162)=12.46$, $p=.001$, $\eta_p^2=.071$, with lower ratings when the second variable was causally active. Most crucially for our purposes, these main effects were qualified by a significant interaction, $F(1,162)=9.77$, $p=.002$, $\eta_p^2=.057$. As shown in Figure S5, we replicated the moderator effect in the “potential moderator” condition (using a new moderator variable, ambient temperature), $p<.001$; in contrast, in the “potential independent cause” condition, presenting the second variable as an independent cause of the same effect did not lead to a drop in the causal ratings of the original cause, $p=.773$. This suggests that the effect we attribute to stability can not be explained by revising beliefs after learning about alternative independent causes, or, more generally, in light of new causally relevant information (given that in both conditions, participants received new information about a causally relevant factor).

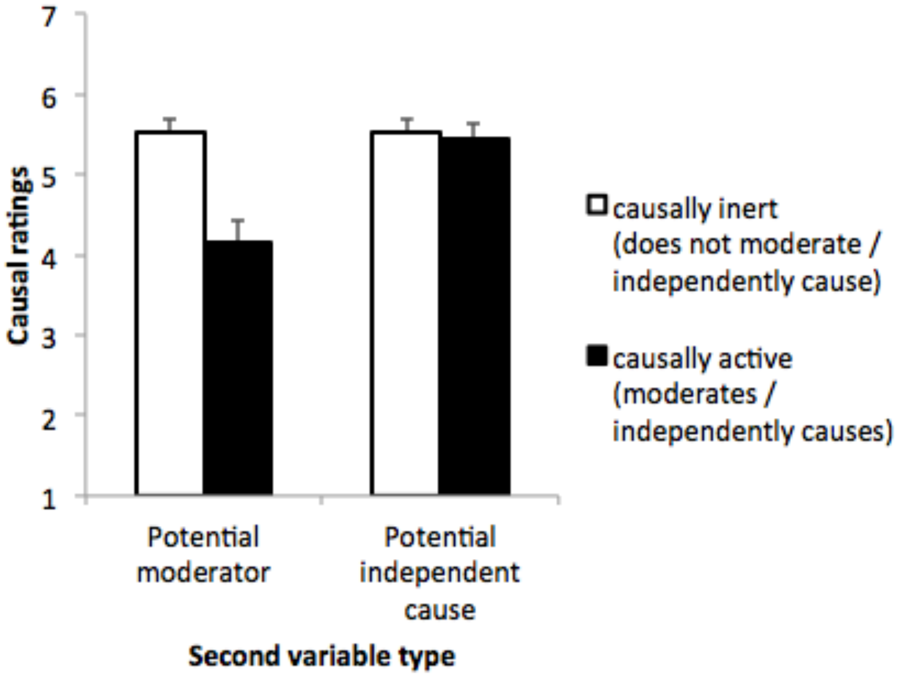


*Figure S5*. Causal ratings of the target variable (eating yona plants) as a function of the second variable type and status. Error bars represent 1 SEM.

**Online supplement E**

In this supplement we offer a clarification of two approaches to averaging causal strength across subpopulations, and we provide the predictions that each approach makes about the outcome of Experiment 2.

There are two ways to compute the average strength of an unqualified causal relationship (e.g., eating yonas → sore antennas): first, participants could calculate the average causal strength by collapsing the two moderator conditions into a single data table and then using it to compute a mental index of causal strength. Second, they could first compute separate mental indices of causal strength for each data table (e.g., for salty water and for fresh water), and only then compute a (weighted or unweighted) mean strength. In Experiment 1, these approaches would yield the same result if people's computation of causal strength is best captured by the ΔP metric (Allan, 1980). However, computing *causal power* (Cheng, 1997) for split tables prior to averaging them would penalize the average causal strength in the moderated condition, introducing an alternative explanation of the pattern observed in Experiment 1.

Fortunately, the frequency manipulation introduced in Experiment 2 allows us to distinguish effects of stability from this alternative account. We used the "first compute then average" approach to calculate average causal strength based on the covariation tables provided to participants, employing either ΔP or causal power metrics, and using either a weighted or unweighted procedure for averaging, where the weighting was determined by the proportion of the population for which each moderator variable value applied. All of these ways of computing strength predict either no effect of moderator (using weighted average ΔP), or that the magnitude of the effect should vary across frequency conditions (see Figure S6 for the resulting distributions of causal strength across frequency conditions). Comparing Figure S6 with Figure 4 (showing the results obtained in Experiment 2), we see that these alternative proposals generate qualitatively different predictions from the pattern of results that was actually observed.

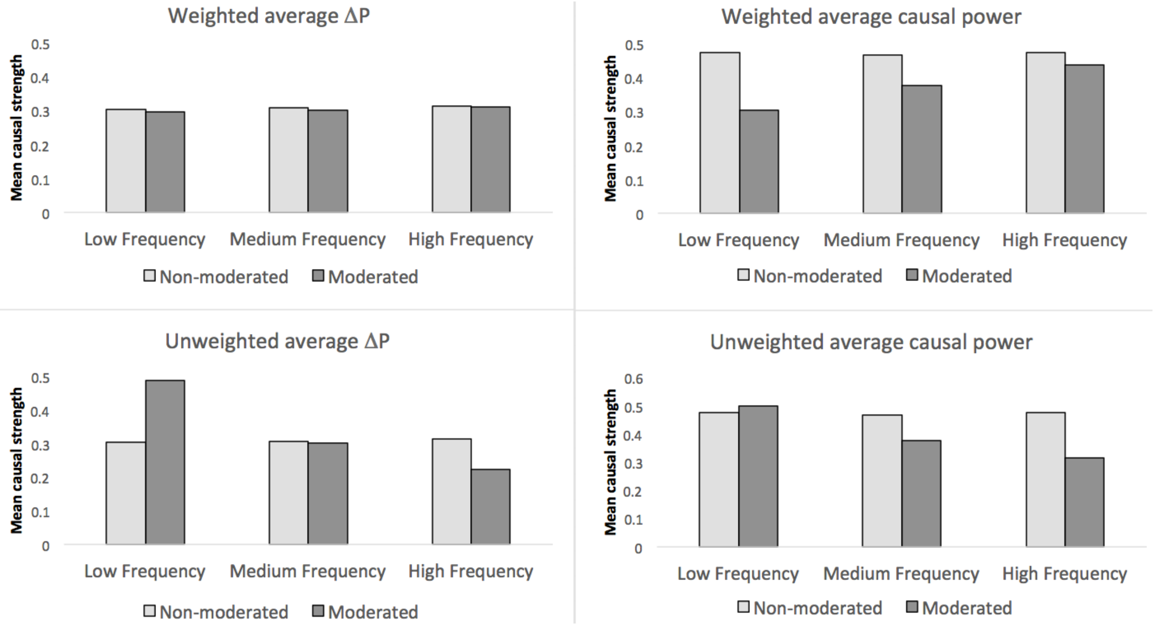


*Figure S6.* Average strength of causal relationships in non-moderated and moderated conditions in Experiment 2 as a function of moderator frequency, computed as either a weighted or unweighted mean of causal strength values calculated for two subpopulations (either using a ΔP or causal power).

**Online Supplement F**

To address the possibility that in Experiment 2 actual scope (frequency) did boost ratings in the moderated condition, but that across frequency conditions this effect was canceled out by decreases in the causal strength of the relationship that held in one subgroup (e.g. salty water), we conducted a new experiment in which we held the causal strength fixed in that subgroup (rather than in the overall population, as we did in Experiment 2), effectively eliminating causal strength as a possible confound within the target subgroup. The strength in the target subgroup was set to $\Delta$P=.72-.74[8]. The rest of the materials, design and procedure were the same as in Experiment 2[9].

If, in Experiment 2, frequency boosted ratings for the moderated relationship but this effect was cancelled out by decreasing causal strength in one subgroup, we should expect the penalty for moderation to weaken across frequency conditions in the modified experiment. However, this is not what we found. Although on both structure and strength measures there was a significant moderator x frequency interaction, $F_{\text{stru}}(2,408)=3.02$, $p=.050$, $\eta_p^2=.015$; $F_{\text{stre}}(2,408)=4.70$, $p=.010$, $\eta_p^2=.022$, the interaction pattern contradicted the alternative explanation of our findings: instead of the effect of moderator

[8] Because the causal strength in the other subgroup had to be near zero, the overall causal strength of the relationship inevitably had to vary across frequency conditions ($\Delta$P=.50-.52, $\Delta$P=.36-.37, and $\Delta$P=.23 for low, medium, and high, respectively).

[9] With the exception that all the
main causal/explanatory questions were qualified (i.e. asked separately for each moderator subgroup); the causal structure and strength ratings were collected both in qualified and unqualified form, as in the Experiment 2. Only the relevant analyses on the unqualified ratings are reported in this Supplement; Online Supplement A contains references to the qualified dataset.

weakening as moderator frequency went up, it showed the opposite trend. Moreover, the interaction effect was driven by the non-moderator condition: while ratings in the moderator condition remained approximately constant across frequency conditions, ratings in the non-moderator condition increased from the low frequency condition to the medium and high frequency conditions. Finally, the main effect of moderator replicated on both measures ($F_{\mathrm{stru}}(1,408)=41.88$, $p<.001$, $\eta_{\mathrm{p}}^{2}=.093$; $F_{\mathrm{stre}}(1,408)=44.04$, $p<.001$, $\eta_{\mathrm{p}}^{2}=.097$), with higher ratings of the non-moderated relationship than the moderated relationship. Overall the results of this study provide further evidence that the moderator effect is not driven by variations in actual scope.

## 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

Cheng, P. W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. *The Psychology of Learning and Motivation, 30,* 215-264. https://doi.org/10.1016/S0079-7421(08)60298-4

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 and J.-A. Meyer (Eds.), *Comparative Approaches to Cognitive Science.* Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545-567. https://doi.org/10.1037/0022-3514.58.4.545

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40,* 83-120. https://doi.org/10.1016/0010-0277(91)90047-8

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365-382. https://doi.org/10.1037/0033-295X.99.2.365

Cheng, P. W., Park, J., Yarlas, A. S., & Holyoak, K. J. (1996). A causal-power theory of focal sets. *The Psychology of Learning and Motivation, 34,* 313-355. https://doi.org/10.1016/S0079-7421(08)60564-2

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

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

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