Revision of Simple Causal Hypotheses: Inferring Interaction Across Multiple Contexts

Mimi Liljeholm (mlil@ucla.edu)
Department of Psychology, UCLA

Patricia Cheng (cheng@psych.ucla.edu)
Department of Psychology, UCLA

Beatrice Leung (bealeung@ucla.edu)
Department of Psychology, UCLA

Abstract

Two competing psychological models of causal strength estimation make different predictions regarding when simple causal hypotheses will be rejected in favor of more complex ones. Two experiments tested these predictions, employing a novel method for indirectly assessing perceived causal strength. In both experiments, the task required a judgment regarding the existence of an interaction between a candidate cause and unobserved background causes. Results indicate that reasoners revise simple causal hypotheses based on the mental construct of *causal power* (Cheng, 1997).

Keywords: Causal power; interaction; indirect assessment; coherent generalization

Inferring an Interaction Across Contexts

The psychological debate about human causal learning has focused on the distinction between covariation and causation, and has been carried out in the context of two models that both yield estimates of causal strength: The ΔP rule (e.g., Jenkins & Ward, 1965), which strictly assesses covariation, and the power-PC theory (Cheng, 1997), according to which reasoners make *a priori* assumptions that allow them to explain covariation by unobservable causal powers in the distal world. Currently, there is evidence for and against both approaches. Specifically, when causal power is held constant while ΔP varies, ratings of causal strength vary ordinally as predicted by ΔP; in contrast, when ΔP is held constant and causal power varies, ratings vary ordinally as predicted by causal power (e.g., Buehner and Cheng, 1997). In other words, strength estimates deviated from both measures.

Buehner, Cheng, and Clifford (2003) found that support for the two opposing measures differed depending on the causal-strength question asked. One might argue that the differences were due to biases in the explicit verbal-assessment questions. In the current experiments, to avoid potential demand characteristics of the strength questions, we employed an indirect measure of perceived causal strength based on reasoners' decision to reject a simple causal hypothesis in favor of a more complex one. Presumably, if the strength of candidate cause *c* is perceived to change across contexts, the reasoner may infer that *c* interacts with unobserved background causes occurring in those contexts. We demonstrate our novel method using the scenarios illustrated in Figure 1.

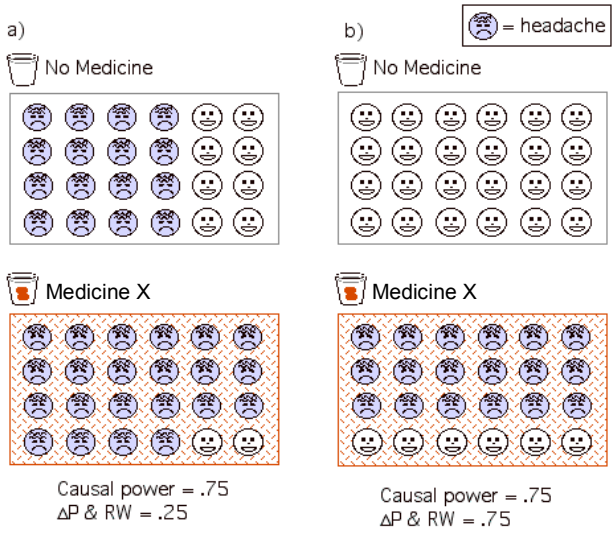

Figure 1: Two hypothetical scenarios across which causal power is constant while ΔP varies.

Imagine that you are presented with data from two studies that both test the influence of an allergy medicine (Medicine X) on headache (a possible side effect). In each study, allergy patients are randomly assigned to two groups: a treatment group and a no-treatment (i.e., control) group. Headache (indicated by a frowning face, as opposed to a smiling face) occurs with different relative frequencies across the two control groups (top panels in the figure), as well as across the two treatment groups (bottom panels). What is your best bet, based on the results from both studies, on whether Medicine X interacts with unobserved background causes that vary across the two studies? As mentioned, if one perceives a "change" in the results across the two studies one might infer the existence of such an interaction. But what constitutes a "change"? Equivalently, what is assumed to be invariant, and hence to generalize, across contexts?

ΔP and Causal Power: Two Psychological Accounts of Causal-Strength Estimation

According to a well-established model of covariation, the *ΔP rule* (Jenkins & Ward, 1965) reasoners contrast the probability of a target effect *e* in the presence of a candidate cause *c,* P(*e*|*c*), and the probability of *e* in the absence of *c,* P(*e*|~*c*),:

$$\Delta P_c = P(e|c) - P(e|\sim c) \qquad (1).$$

These probabilities are directly estimable by the observable relative frequencies of the relevant events. Under a set of conditions, the ΔP rule is equivalent to Rescorla and Wagner's model (1972; see Danks, 2003) when learning is at asymptote; both accounts have been adopted to model causal strength (e.g., Spellman, 1996). For all conditions in our experiments, these accounts make identical ordinal predictions.

Across the studies in Figure 1, from left to right, *ΔP* of the treatment increases from .25 to .75. Therefore, if judgments are based on *ΔP*, people should perceive a change in the results across treatments and, accordingly, infer that Medicine X interacts with background causes.

In contrast, according to the *causal power theory of the probabilistic contrast model* (Cheng, 1997; the *power PC* theory for short), the reasoner explains observable probabilistic contrast (e.g., covariation) by postulating unobservable causal relations in the distal world. The *causal power* (i.e., magnitude) of these relations are indirectly estimable under a set of generic causal assumptions. For example, as default assumptions that can be revised in light of evidence, candidate *c* influences *e* independently of background causes of *e*, and the latter causes do not prevent *e*. From that set of assumptions, if in addition background causes are believed to occur independently of *c*, then it follows that when $\Delta P \geq 0$, the *generative power* of *c* with respect to *e* can be estimated as follows:

$$\textit{generative power of } c = \frac{P(e \mid c) - P(e \mid \sim c)}{1 - P(e \mid \sim c)} \qquad (2)$$

The preventive analog of this equation applies when $\Delta P \leq 0$ (see Cheng, 1997, for derivations).

Across the two studies depicted in Figure 1, the causal power of the treatment remains at 0.75 according to Equation 2. Therefore, if "change" is defined by causal power values, people should not infer that Medicine X interacts with background causes.

Generalization Across Contexts

The assumption of the independent influence of *c* and the background causes is critical for coherent generalization (Liljeholm & Cheng, in press). If the influence of a cause depends on (i.e., the cause interacts with) context-specific and potentially unobserved background causes, knowledge acquired in one context will not apply in another. Note that if two or more causes *independently influence* an effect *e*, then when they are jointly present, *each operates on e as if the other causes were not there*. To illustrate this concept for the two studies in Figure 1, consider administering Medicine X to patients in both control groups (top panels). Assume that the medicine causes headache with a probability of 3/4 in *each and every one* of the 24 patients in each control group, operating in the same way regardless of which figure it is in (i.e., as if the background causes were not there). The resulting frequencies of headache would be consistent with what appears in the respective treatment groups (the bottom panels), in accordance with the definition of independence in probability theory. In particular, for the treatment group in Figure 1a, the probability of headache being caused independently by both the medicine and the background should be 1/2, the product of the probabilities of the constituent events: the probability that headache is caused by the medicine (3/4) and by the background (2/3). It follows that 11/12 is the probability of headache being caused by the medicine or the background (the sum of the constituent probabilities minus their product), matching the observed outcome in that panel.

In contrast, the *ΔP* model makes the following anomalous and counterintuitive prediction. Let us return to Figure 1b (right panels), in which Medicine X has a *ΔP* value of .75. Now, consider testing this medicine on the patient groups in Figure 1a. Not only does the covariational approach conclude that the medicine interacts with the background causes in view of the results for these additional patients, it concludes that it is doomed to do so: given the top left panel (the control group for that study), it is *impossible* for the medicine to have a *ΔP* value of .75 -- one therefore need not conduct the study to find out.

Previous Tests of Causal Strength Estimation

If *ΔP* makes such incoherent and anomalous predictions, the question remains: Why do causal strength ratings vary with *ΔP* when causal power is held constant (e.g., Buehner & Cheng, 1997), contradicting the causal power view? There is an explanation besides ΔP for such ratings. Let us return to Figure 1, but now replace Medicine X in Figure 1b with another candidate cause (e.g., Medicine Y). If you are asked to rate the causal influence of each medicine on headache, would your ratings differ across the two medicines? Presumably, the answer would depend on the question.

For example, if you are asked, "how likely is it that Medicine X (or Y) causes headache?", your judgment might incorporate your level of confidence due to sample size. Note that whereas the actual sample size remains constant across the two scenarios illustrated in Figure 1, the *virtual sample size* – defined as the estimated number of trials on which the production of headache by the candidate cause can be unambiguously evaluated – is greater for the second scenario (i.e., 24) than for the first (i.e., 8). Consequently, if your rating reflected your level of confidence due to sample size, it should be lower for the first scenario than for the second. Griffiths and Tenenbaum's (2005) Bayesian *causal support* model is a formal answer to the "how likely"

question based on either ΔP or causal power; the causal-power variant explains the puzzling deviations from both ΔP and causal power observed by Buehner and Cheng (1997).

In previous tests of the power-PC theory, some experimenters who intended to measure causal strength have asked questions that are ambiguous with respect to the separation of strength and confidence. For example, in Buehner and Cheng's (1997) generative condition participants were asked to rate "how strongly they thought" a particular type of rays cause mutation in a strain of virus, on a causal-strength scale from 0 to 100. Note that "strongly" can qualify either verb, "thought" or "cause"; that is, "how strongly they thought…?" and "how strongly … beta rays cause mutation…?" are two possible interpretations of the question.

Alternatively, a question intended to measure strength may be interpreted by some subjects as pertaining to the degree that a candidate cause increases or decreases the probability of an effect in the learning context, in which case the answer should correspond to ΔP. In sum, in view of the ambiguity in the materials of previous studies, observing answers that deviated from the predictions of causal power does not refute the model.

Although it is possible to ask less ambiguous questions, clarification may bias participants towards a particular measure and therefore fail to reveal anything about the spontaneous use of that measure. In contrast, our interaction query allows one to indirectly assess participants' estimates of causal strength without biasing them.

Experiment 1: A structural decision about a conjunctive causal link based on multiple contexts

It is only by comparing causal-strength estimates of a single candidate cause across distinct contexts that a reasoner can assess if the independent causal influence assumption is violated or confirmed. A violation (i.e., the candidate's causal strength changes across contexts) provides a signal for seeking a more complex explanation. Experiment 2 addresses the issue: What is the criterion for deciding that a simple causal hypothesis needs revision? The answer rests on the concept of invariance across contexts.

Method Thirty-eight undergraduates from the University of California, Los Angeles, participated to obtain course credit in an introductory psychology course. They were randomly assigned to each of two groups. In both groups, participants were presented with two hypothetical studies similar to those in Figure 1 that tested the influence of a fictitious allergy medicine on headache. For one group, the *Power-Constant* Group, causal power was held constant across the two studies while ΔP varied (as in Figure 1). For the second group, the *Power-Varying* Group, both causal power and ΔP varied across the two studies. The relative frequencies of the events and the model values for each hypothetical study for both groups are listed in Table 1.

The primary measure was a judgment based on *both* hypothetical studies about whether the medicine interacts with background causes. Specifically, participants were asked, "Based on the results from BOTH experiments, do you think that Medicine X interacts with some factor that varies across experiments, or do you think that the medicine influences the patients in different experiments in the same way?" We did not specify what "the same way" means. Participants were given three answer options: "Yes", "No" and "Can't Tell". If judgments about an interaction are based on ΔP, one would expect as many participants to infer an interaction in the Power-Constant Group as in the Power-Varying Group. In contrast, if such judgments are based on causal power, more participants should infer an interaction in the Power-Varying Group than in the Power-Constant Group. A secondary measure was a rating of the causal strength of the medicine, based on each individual study. Specifically, after viewing each study participants were asked: "Based *only* on the results from this experiment, please rate how strongly medicine X causes headache." The ratings were made on a 100-point scale ranging from 0 ("never causes headache") to 100 ("always causes headache"). This rating scale was intentionally ambiguous for two reasons: First, to avoid biasing participants towards any particular model and, second, to explore whether inferences about an interaction would vary with causal power even when causal ratings reflected an alternative interpretation of the strength question.

Table 1: Experiment 1 - Relative frequencies of headache and model values for each hypothetical study (Studies a & b) in 2 Groups: Power-Constant (P-C) and Power-Varying (P-V).

	Frequencies of Headache				Model Predictions			
	Study a		Study b		Power		ΔP & RW	
	eInoX	eIX	eInoX	eIX	Studya	Studyb	Studya	Studyb
P-C	24/64	64/64	0/64	64/64	1.0	1.0	.63	1.0
P-V	0/64	40/64	0/64	64/64	.63	1.0	.63	1.0

The stimuli and questions were presented on paper. First, a cover story informed participants that a pharmaceutical company was investigating if an allergy medicine (Medicine X) might produce headache as a side effect, and that they would be shown the results from two experiments that tested the influence of this medicines. Participants were further instructed that the two experiments had been conducted in different labs and that, therefore, the number of allergy patients who had a headache without having received any medicine may vary across studies. Below the cover story appeared data from Study a, the first study. The medicine and headache were graphically depicted as in Figure 1 except that patients with headaches were always grouped by column on the left side of a panel. At the bottom of the page, participants were presented with the causal-strength question and rating scale described above. Study b was presented in the same format on a second page. Finally, after viewing and providing strength ratings for both studies, participants were asked the interaction query.

Results Our results demonstrate that revisions of a simple causal hypothesis are in accord with causal power. Whereas less than 1/3 of the participants in the Power-constant Group (6 out of 19) responded "yes" to the interaction query, more than 3/4 (15 out of 19) in the Power-varying Group did so, χ^2 (1, N= 38) = 8.62, p<0.005.

In contrast, the ratings of causal strength varied across studies in both groups, as predicted by ΔP. A Group (2) X Study (2) analysis of variance was performed on the strength ratings with Study as a within-subjects variable.

The mean strength ratings were lower for Study a (mean = 73.42) than for Study b (mean = 98.68) resulting in a main effect of Study, F(1, 36) = 214.27, MSE = 56.59, p <0.001. There was also a main effect of Group: mean ratings were lower in the Power-Varying Group (mean = 81.97) than in the Power-Constant Group (mean = 90.13), F(1, 36)=14.61, MSE= 86.56, p<0.001. In addition, there was a Group by Study interaction, such that the difference between mean ratings for the two studies was greater in the Power-Varying Group (mean difference= 35) than in the Power-Constant Group (mean difference= 15.5), F(1, 36)=31.83, p<0.001. Planned comparisons revealed that in the Power-varying Group, the mean rating for Study a (mean=64.5, SD=6.38) was lower than that for Study b (mean=99.5, SD=2.29), t(18)=23.83, p<0.001. Likewise in the Power-constant Group, the mean rating for Study a (mean=82.4, SD=15.0) was lower than that for Study b (mean=97.9, SD=3.84), t(18)= 4.97, p<0.001.

Interestingly, for the Power-constant Group, if causal ratings for the 6 participants who responded "yes" to the interaction query are eliminated, for the rest of the group the mean of the ratings for Study a is still lower than that for Study b (mean difference=11.46, t(12)=3.56, p<0.005). In other words, even those participants whose strength ratings varied across data sets, as ordinally predicted by ΔP, did not infer an interaction when causal power remained constant.

Experiment 2: A control for the influence of P(e|c)

In the previous experiment, P($e|c$), the probability of effect e in the presence of cause c, was 1.0 for both studies in the Power-Constant Group but varied across studies in the Power-Varying Group. It is possible that participants simply based their judgments about whether the influence of c differs across data sets on how often e occurs in the presence of c. In Experiment 2 causal power is instead held constant at 0.75 in the Power-Constant Group and increases from 0.25 to 0.75 in the Power-Varying Group. Meanwhile, P($e|c$) varies across hypothetical studies in both groups. Our two experiments used the same method, except that participants in Experiment 2 were presented with 4 hypothetical studies rather than 2.

Method As in Experiment 1, UCLA undergraduates in an introductory psychology course (N=40) were randomly assigned to two groups: the *Power-Constant* Group and the *Power-Varying* Group. In the Power-Constant Group, ΔP increased across studies from .19 to .75 while causal power remained constant at .75 across studies. In the Power-Varying Group, both causal power and ΔP increased across studies from .19 to .75. For both groups and for each hypothetical study, Table 2 lists the relative event frequencies and model values. The four studies, Studies a, b, c and d, occurred respectively in two counterbalanced orders [1st, 3rd, 2nd, 4th & 4th, 2nd, 3rd, 1st].

Table 2: Experiment 2 - Relative frequencies of headache and model values for each hypothetical study (a-d) in 2 Groups: Power-Constant (P-C) and Power-Varying (P-V).

	Relative Frequencies of Headache							
	Study a		Study b		Study c		Study d	
	eInoX	eIX	eInoX	eIX	eInoX	eIX	eInoX	eIX
P-C	48/64	60/64	32/64	56/64	16/64	52/64	0/64	48/64
P-V	0/64	12/64	0/64	24/64	0/64	36/64	0/64	48/64

	Model Values							
	Causal Power				ΔP & RW			
Study	a	b	c	d	a	b	c	d
P-C	.75	.75	.75	.75	.19	.38	.56	.75
P-V	.19	.38	.56	.75	.19	.38	.56	.75

The materials and task were identical to those in Experiment 1 with two exceptions. First, the stimuli were presented on the computer. Second, the cover story was modified to indicate that participants would be shown the results from four, rather than two, tests of the medicine. Participants were presented with 4 consecutive screens, each of which showed a particular group of patients before and after they received the mineral.

Results Our results corroborate those of Experiment 1. Whereas almost all participants in the Power-Varying Group (18 out of 20) responded "yes" to the interaction query, only about 1/3 (7 out of 20) in the Power-Constant Group did so, χ^2(1, N=40)=12.9, p<0.001.

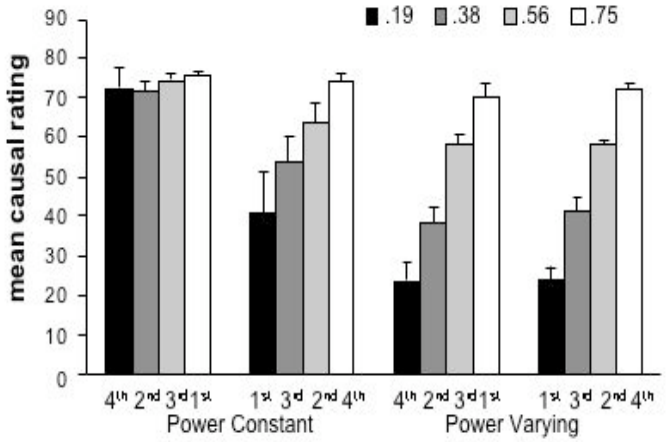

Figure 2: Mean causal strength ratings from Experiment 2 as a function ΔP and Group. Numerical labels on the x-axis indicate the order in which the hypothetical studies listed from left to right occurred.

Figure 2 shows the mean causal ratings as a function of ΔP in both groups and for each counterbalancing order. A

Study (4) x Group (2) x Order (2) analysis of variance on the ratings was performed on the causal strength ratings with Study as a within-subject variable. Both expected main effects, of Study and Group, were significant, as was a Study x Group interaction. In addition, Order interacted with both of the other two variables.

The mean of the strength ratings for the Power-Constant Group was higher than that for the Power-Varying Group, $F(1, 36)=36.71$, $p<0.001$. But, as in Experiment 1, contrary to the power values, causal-strength ratings varied across studies for both Groups, resulting in a main effect of Study, $F(3, 108)=49.73$, $p<0.001$. Simple effects analysis revealed that the difference between the two extreme studies was highly significant in both Groups, $t(19)>2.60$, $p<0.02$.

At the same time, contrary to ΔP, the differences between the studies were greater in the Power-Varying Group than in the Power-Constant Group, resulting in a Group by Study interaction, $F(3, 108)=12.49$, $p<0.001$. Specifically, the difference between the two extreme studies (i.e., Studies a and d) was greater in the Power-Varying Group (mean difference=49.75) than in the Power-Constant Group (mean difference=14.6). Interestingly, in the Power-Constant Group, but not in the Power-Varying Group, the difference between ratings for different studies depended on the order in which those studies were presented, resulting in a Group x Study x Order interaction, $F(3,108)=3.48$, $p=0.018$, as well as a marginally significant Study x Order interaction, $F(3,108)=2.59$, $p=0.057$. Simple effect analyses revealed that, in the Power-Constant Group, mean ratings for Study a, the study with the smallest ΔP, were significantly lower than mean ratings for Study d, the study with the largest ΔP when both studies were presented in the first position, $t(18)=3.42$, $p<0.005$, but not when they were presented in the fourth position, $t(18)=0.3$, $p=0.77$. In contrast, in the Power-Varying Group, the mean ratings for Study a were significantly lower than those for Study d, regardless of the counterbalancing order, $t(18)>9.6$, $p<0.001$.

Note that, if the increasing causal-strength ratings across studies reflected the influence of ΔP, then one would have to explain, not only why the influence of ΔP was weaker in the Power-Constant Group, but also why this influence changed depending on order only in the Power-Constant Group. Thus, ΔP can be eliminated as explanation even for causal-strength ratings.

An alternative explanation is that for the Power-Constant Group, the medicine is perceived as operating the same way across studies, and therefore data from additional studies can be accumulated. When the studies with extreme values of ΔP were presented in the first position, participants in this group would have lower confidence for the study with the smallest ΔP due to its small virtual sample size, compared to the study with the largest ΔP. However, when these studies were presented in the last position, the studies with the smallest and largest ΔPs would be added to the cumulative samples from previous studies and thus become comparable across orderings. In contrast, in the Power-Varying Group, because the medicine interacts with the background causes, data from different studies cannot be collapsed, and Order therefore should have no effect. Presumably, for this group, the differences in causal ratings across studies are due to variations in causal power rather than virtual sample size.

General Discussion

Both experiments indicate that people tacitly adopt generic assumptions regarding unobservable causal events so that coherent generalization across contexts is possible. Decisions regarding an interaction between candidate and background causes varied according to causal power rather than ΔP. The same pattern of results was obtained with variations of our experimental materials: scenarios involving a between-subject rather than before-and-after design, a sequential rather than simultaneous presentation of trials, and other values of causal power, ΔP, and scenario sample size (Liljeholm & Cheng, in press).

To address some potential concerns about the experiments presented here, one of these replication experiments deserves a more detailed discussion. One concern is that, in both Experiment 1 and 2, ΔP varied across the hypothetical studies in both the Power-Constant Group and the Power-Varying Group. It is possible that participants simply based their answer on whatever measure was held constant. Moreover, in the current experiments the unobserved, background causes never produced headache in the absence of Medicine X for the Power-Varying Groups (i.e., the base rate of e always equaled 0) but did so in one or more hypothetical study for the Power-Constant Groups. Thus, it is possible that our results reflect these differences between groups rather than assessments of causal power per se. Finally, it may be argued that our interaction query was too complex for participants to understand, making their answers difficult to interpret.

Liljeholm and Cheng (in press) reported experiments that addressed all of these concerns, and obtained results consistent with those reported here. In their Experiment 1, one group of participants, the ΔP-Constant Group, was presented with the two hypothetical studies illustrated in Figure 3. Across the studies in this figure, from left to right, causal power increases from .25 to .75 as the base rate of e increases, while ΔP of the treatment remains constant at 0.25, thus allowing participants who base their response on a lack of change to make use of ΔP. To allow a less complex question, in the second study, allergy patients receive *both* Medicines X and Y (see bottom panel in Figure 3b). Rather than asking about the existence of an interaction, Liljeholm and Cheng (Experiment 1; in press) asked for judgments, based on both studies, regarding whether Medicine Y was a cause of headache. As with the inference of an interaction, if one perceived a "change" in the results across treatments (i.e., Medicine X in one study and both medicines in the other) this change should be attributed to the introduction of Medicine Y in the second study. The other group of participants in the experiment, a

Power-Constant Group, were presented with the two hypothetical scenarios illustrated in Figure 1 above, except that allergy patients in the second scenario (Figure 1b) received both medicines X and Y. Unlike in Figures 1 and 3, individual trials in each scenario (corresponding to individual faces in the figures) were presented to participants in random sequential order, except that trials in corresponding positions in the control and treatment panels

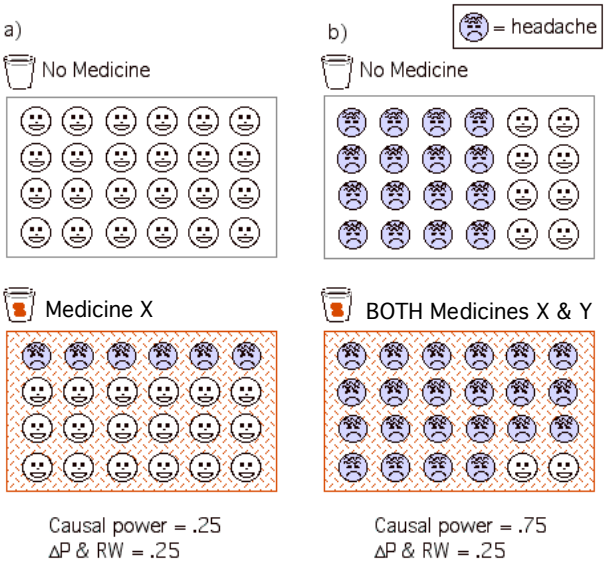

Figure 3: Two hypothetical scenarios across which causal power varies while ΔP is constant.

were paired in a before-and-after design.

In summary, in Liljeholm and Cheng's (in press) experiment, changes in the base rate of *e* were held constant, and the manipulations of ΔP and causal power were symmetrical across groups. In addition, the dependent measure was a query about the existence of a simple causal link rather than an interaction. Finally, trials were presented in a randomized and sequential format. The pattern of results was identical to those reported here: whereas 4/5 of the ΔP-Constant Group responded that Medicine Y was causal, less than 1/3 of the Power-Constant Group did so.

Recently, Griffith and Tenenbaum (2005) have argued that, while previous psychological models address the estimated strength of a causal relationship, human judgments often reflect a structural decision about whether or not a causal relationship exists. Their Bayesian *causal support* model is a normative measure of the amount of evidence provided by a sample in favor of the existence of an elemental causal relation. This measure and χ^2 are highly correlated (Griffiths & Tenenbaum, 2005); for example, both increase with sample size, other things being equal. Is causal support something that reasoners carry from one context into another? The clear answer is "no". In Experiment 2, causal support values increased from 3.9 to 44.8 across the four studies in the Power-Constant Group, and from 5.4 to 44.8 in the Power-Varying Group. Thus, if judgments about whether Medicine X interacts with background causes had been based on a change in causal support values across studies, more participants should have inferred an interaction in the Power-Constant Group than in the Power-Varying Group, opposite to the pattern of results observed here (also see Liljeholm & Cheng, in press, Experiment 2). Note, however, that our results do not pose a problem for the Bayesian approach per se. Bayesian models appropriate for the structural decision in our experiments can be developed (e.g., see Jaynes, 2003) and indeed have been developed (T. Griffiths, personal communication, 2006).

Taken together, results reported by Liljeholm and Cheng (in press) and here clearly indicate that causal power is the feature people carry from one context into another.

References

Buehner, M. J., & Cheng, P.W. (1997). Causal induction: The power PC theory versus the Rescorla–Wagner theory. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 55–60). Mahwah, NJ: Erlbaum.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 29(6), 1119-1140.

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

Liljeholm, M & Cheng, P. W., (in press). When is a cause the 'same': Coherent Generalization across contexts. *Psychological Science*

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*. 47(2), 109-121

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

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jenkins, H.M., & Ward, W.C. (1965) Judgment of contingency between responses and outcomes, *Psychological Monographs*: *General and Applied*, 79 (1, Whole No. 594).

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century Crofts.

Spellman, B. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds) *Advances in the psychology of learning and motivation, Vol. 34: Causal Learning* (pp. 167-206). New York, NY: Academic Press.