Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 1, 157–172

0278-7393/09/$12.00 DOI: 10.1037/a0013972

The Influence of Virtual Sample Size on Confidence and Causal-Strength Judgments

Mimi Liljeholm and Patricia W. Cheng
University of California, Los Angeles

The authors investigated whether confidence in causal judgments varies with *virtual sample size*—the frequency of cases in which the outcome is (a) absent before the introduction of a generative cause or (b) present before the introduction of a preventive cause. Participants were asked to evaluate the influence of various candidate causes on an outcome as well as to rate their confidence in those judgments. They were presented with information on the relative frequencies of the outcome given the presence and absence of various candidate causes. These relative frequencies, sample size, and the direction of the causal influence (generative vs. preventive) were manipulated. It was found that both virtual and actual sample size affected confidence. Further, confidence affected estimates of strength, but confidence and strength are dissociable. The results enable a consistent explanation of the puzzling previous finding that observed causal-strength ratings often deviated from the predictions of both of the 2 dominant models of causal strength.

Keywords: causal strength, confidence, virtual sample size

Suppose you are testing whether a particular mineral in an allergy medicine causes headache as a side effect. Consider the hypothetical scenario illustrated in Figure 1A, where headache occurs on 48 of 64 trials when the mineral is absent and on 64 of 64 trials when it is present. Each face in the figure represents a trial. Assume that alternative causes of headache occur independently of the mineral (i.e., they occur just as often in the presence as in the absence of the mineral). Now imagine a slightly different scenario, in which headache occurs on 16 of 64 trials when the mineral is absent and on 64 of 64 trials when the mineral is present, as illustrated in Figure 1B. What would be your impression of the strength of the mineral's influence on headache in each of these scenarios?

On the one hand, you might simply consider the number of trials out of the total on which the mineral made a difference. In this case, for the first scenario, in which the mineral makes a difference on 16 of 64 trials, your estimate of the mineral's influence should be low (i.e., 16/64 = 25%) relative to the second scenario, in which the mineral makes a difference on 48 of 64 trials (i.e., 48/64 = 75%). This approach corresponds to a well-established covariational model of causal strength, the Δ*P rule* (Jenkins & Ward, 1965), according to which the reasoner contrasts the probability of a target effect *e* (e.g., headache) in the presence of the candidate cause *c* (e.g., Mineral X), P(*e*|*c*), with the probability of *e* in the absence of *c*, P(*e*|∼*c*):

$$\Delta P_c = P(e|c) - P(e|{\sim}c) \tag{1}$$

According to this model, the estimates of the minerals causal strength should be lower for the first scenario, .25, than for the second, .75.

On the other hand, you might consider the possible direction of change in the effect: Only on trials in which there is no headache to begin with can one assess the mineral's ability to cause a headache. Consider the simple case in which each scenario concerns patients before and after being administered the mineral. For those 16 trials in the first scenario (Figure 1A) and 48 trials in the second scenario (Figure 1B), the mineral causes headache every time (i.e., 16/16 = 100% and 48/48 = 100% for the two figures, respectively). This analysis illustrates an application of the power-PC theory (Cheng, 1997; Novick & Cheng, 2004).[1] In contrast to purely covariational models, this causal account posits that the reasoner makes assumptions about unobservable causal relations to explain observable events. For example, one default causal assumption (that can be overridden if refuting evidence becomes available) is that the candidate cause and alternative causes (e.g., the mineral and muscle tension in the headache example) influence the effect independently.

It can be shown that under the assumptions specified by the power-PC theory, if alternative causes of *e* are believed to occur

Mimi Liljeholm and Patricia W. Cheng, Department of Psychology, University of California, Los Angeles.

Experiment 1 was part of a master's thesis submitted by Mimi Liljeholm to the Department of Psychology, University of California, Los Angeles, under the supervision of Patricia W. Cheng, and was presented at the Psychonomic Society 46th Annual Meeting, Toronto, Ontario, Canada, November 2005. The research reported in the article was supported by National Institutes of Health Grant MH64810 to Patricia W. Cheng. We thank Keith Holyoak, David Lagnado, Pierre Mercier, and Edward Wisniewski for their helpful comments on the article.

Correspondence concerning this article should be addressed to Mimi Liljeholm, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563. E-mail: mlil@ucla.edu

[1] The theory applies to the before-and-after design as long as subjects are willing to make the "no confounding" assumption (i.e., the assumption that alternative causes occur independent of the candidate cause) across the two time frames.

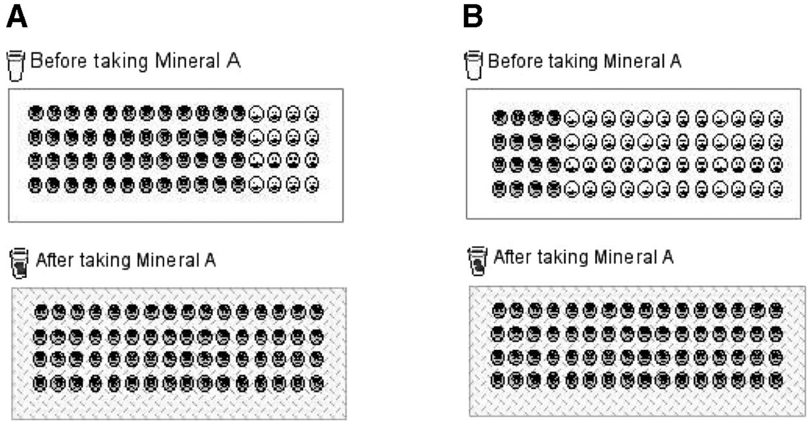

Figure 1. Two hypothetical scenarios (A and B) with different relative frequencies of headache given the absence and presence of Mineral A. Gray faces represent the presence of headache and the background pattern represents the presence of Mineral A.

independently of c, then when $\Delta P > 0$, the *generative causal power* of c with respect to e can be estimated as follows:

$$\text{generative causal power of } c = \frac{P(e|c) - P(e|\sim c)}{P(\sim e|\sim c)} \qquad (2)$$

The generative causal power of c with respect to e is the unobservable theoretical probability with which c produces e (Cartwright, 1989).[2] Similarly, under the same set of assumptions but when $\Delta P < 0$, the *preventive power* of c with respect to e—the unobservable probability with which c prevents e—can be estimated as follows:

$$\text{preventive power of } c = \frac{P(e|\sim c) - P(e|c)}{P(e|\sim c)} \qquad (3)$$

Note that in Equations 2 and 3, the estimation of causal strength is a function of the proportion of cases in which the outcome due to the cause—e which is for generative candidates and $\sim e$ for preventive candidates—does not already occur in the absence of that cause (see the denominators, $P(\sim e|\sim c)$ and $P(e|\sim c)$, respectively, on the right-hand side of Equations 2 and 3). For the set of cases specified by this proportion in both Figures 1A and 1B (Equation 2 for both figures), headache always occurred in the presence of the mineral. Consequently, causal power according to Equation 2 remains constant at 1.0 across the two scenarios.

Deviations of Observed Causal-Strength Judgments From Both Theoretical Approaches

Previous studies have shown that when causal power was constant while ΔP was manipulated, as is the case across the two scenarios illustrated in Figures 1A and 1B, observed causal-strength ratings tended to vary ordinally with ΔP (e.g., Buehner, Cheng, & Clifford, 2003, Experiment 1; Lober & Shanks, 2000). However, when ΔP was instead held constant while power was manipulated, causal-strength ratings were ordinally predicted by power (e.g., Buehner & Cheng, 1997; Lober & Shanks, 2000). Our goal in the present article is to test plausible explanations for this puzzling departure of the observed mean strength ratings from both of the two dominant models of causal strength. We review two hypotheses and present two experiments testing implications of the hypotheses.

The Weighted-ΔP Hypothesis

Some researchers have argued that human reasoners weigh some combinations of c and e more heavily than others when evaluating contingencies (e.g., the joint presence of c and e weighing more than their joint absence) and have shown that introducing weights for the different combinations of events and calculating ΔP on the basis of this weighted information improves the model's fit to observed causal judgments (e.g., Anderson & Sheu, 1995). Similarly, Lober and Shanks (2000) proposed that the deviation from both causal power and ΔP predictions can be explained by differentially weighing the two conditional probabilities in the ΔP rule. They reported that assigning a weight of 1 to $P(e|c)$ and a lower weight (0.6) to $P(e|\sim c)$ provided a good fit to their experimental results. They concluded that covariational models provide a better explanation of empirical results than does the power-PC theory.[3]

The Conflation Hypothesis

An alternative hypothesis is that although reasoners may compute causal power to represent causal strength, ambiguity in questions intended to measure strength may result in observed estimates that do not correspond to causal-power values. Returning to the two scenarios illustrated in Figure 1, note that whereas the actual sample size is constant across the scenarios, the *virtual*

[2] Throughout the remainder of the article, we use the term *causal power* to indicate causal strength as defined by the power-PC theory (Cheng, 1997). We shorten *generative causal power* to *generative power* and likewise *preventive causal power* to *preventive power*.

[3] Lober and Shanks (2000) also proposed that Rescorla and Wagner's (1972) model with opposite orderings of unequal beta weights for generative and preventive scenarios can explain the results, but this explanation has been ruled out by subsequent findings (see Buehner et al., 2003). We therefore only evaluate the weighted-ΔP model as a viable covariational account.

sample size[4]—defined as the estimated number of trials on which the production of headache can be unambiguously evaluated—is much smaller for the first scenario (16) than for the second (48). Conversely, for the preventive influence of the mineral, the virtual sample size is symmetrically greater for the first scenario (48) than for the second (16). One might expect, therefore, that confidence in estimates of casual strength might be greater for the first scenario than for the second.

According to the *conflation hypothesis* (Buehner & Cheng, 1997; Buehner et al., 2003), a difference in confidence in estimates of causal strength due to a difference in virtual sample size, combined with a conflation of strength and confidence due to ambiguity in the experimenter's question, explains why causal judgments observed by Buehner and Cheng (1997) deviated from causal power. Deviations due to virtual sample size would be in the direction predicted by ΔP because, for a given value of causal power, when actual sample size is constant as it was in Buehner and Cheng's (1997) experiment, virtual sample size increases with an increase in ΔP (see both Equations 2 and 3).

Our primary goal in the present article is to test the influence of virtual sample size on confidence and estimates of causal strength, predicted by the conflation hypothesis but not by the weighted ΔP hypothesis.[5] As should be clear, the power-PC theory is an account of causal strength and, therefore, makes no predictions about confidence or about a conflation of strength and confidence. However, the theory does specify that to estimate causal strength, one divides the conditional contrast (i.e., ΔP) by the proportion of the actual sample that makes up the virtual sample; the denominators of Equations 2 and 3 correspond to those proportions for generative and preventive causes, respectively. In this sense, the power-PC theory is consistent with an explanation based on the influence of virtual sample size on confidence.

As mentioned, the conflation hypothesis attributes the influence of virtual sample size on causal-strength ratings to ambiguity in the causal-strength question. In Buehner and Cheng's (1997) generative condition, for example, 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 ranging from 0 (*the rays do not cause the virus to mutate at all*) to 100 (*the rays cause the virus to mutate every time*). 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. This ambiguity may have encouraged participants to conflate their strength estimates with confidence in those estimates.

The conflation hypothesis predicts that less ambiguous questions will allow participants to better reveal their representation of causal strength as causal power. Alternatively, perhaps causal-strength estimates and confidence in those estimates are indissociable, so that reasoners have no choice but to give a single combined estimate regardless of the ambiguity of the strength question. A secondary goal is to assess the dissociability of strength estimates and confidence in those estimates. To our knowledge, no previous study has presented evidence for both causal-strength ratings that are unaffected by sample size and confidence ratings that are a direct function of sample size.

A New Method: A Pure Assessment of the Influence of Virtual Sample Size on Confidence

One difficulty with assessing the influence of virtual sample size on confidence in strength estimates is that this variable can be confounded with other relevant variables. For example, as just mentioned, given a constant actual sample size and causal power, virtual sample size is directly proportional to ΔP. The influence of virtual sample size on confidence is therefore unclear. Likewise, given a constant actual sample size and ΔP, virtual sample size is inversely proportional to causal power. The influence of virtual sample size on confidence may therefore be masked by that of causal power. Previous studies assessing the influence of virtual sample size on confidence in strength estimates have not overcome these problems (e.g., Shanks, 2002). Here, to disambiguate the influence of virtual sample size on confidence from any influence of ΔP and to prevent any influence of causal power from masking that of virtual sample size, we introduce a new method for assessing confidence that is inherently completely unbiased by perceived causal strength.

Consider questions regarding causation versus prevention about the top panel of Figure 1A, the control condition in the mineral experiment. The virtual sample size would be apparent from this top panel alone if one is explicitly asked to evaluate whether the drug causes headache: It is simply the number of cases in which headache was not initially present (i.e., 16). Analogously, if one is explicitly asked to evaluate the relief of headache, the virtual sample size would be the number of cases in which headache was initially present (i.e., 48). In this before-and-after design for both causal directions, the actual sample size is 64. Thus, given knowledge about the causal direction to be evaluated (i.e., generative or preventive) in this type of design, both actual and virtual sample size information is available before the drug is administered. Our new confidence measure made use of this availability.

Experiment 1

We presented participants with scenarios such as those illustrated in Figure 1 and first measured confidence levels solely on the basis of information about the control condition, the group of patients before they receive the mineral (hereafter the *before condition*): Participants were asked, given the causal direction to be evaluated, how confident they would feel about any subsequent results from that particular group of patients. For obvious reasons, this *initial confidence* rating is not influenced by the causal strength of the mineral. We then presented the same patient group after it had received the mineral (hereafter the *after condition*), measuring (a) perceived causal strength and (b) confidence about the perceived strength. We refer to the latter as the *final confidence* rating.

The before-and-after design was used to make it absolutely clear to participants that the size of the sample would remain constant across the absence and presence of a particular mineral. We independently manipulated four variables. We manipulated virtual sample size, by changing the base rate of *e,* for two reasons: (a) to

[4] The term *virtual* is used in the sense of "being such in essence or effect though not formally recognized or admitted" (Mish et al., 2003, p. 1397).

[5] Throughout the remainder of the article, the term *confidence* is used to indicate confidence in estimates of causal strength unless otherwise specified.

test the influence of virtual sample size on the initial confidence rating and (b) to assess how well the deviations of causal-strength ratings from causal-power predictions are explained by the initial-confidence ratings. The initial rather than the final measure of confidence is used for this purpose because, whereas the latter is potentially influenced by estimated causal strength, the former cannot possibly be so. We also manipulated causal power to test the influence of this variable on the final confidence rating, to provide converging evidence for a conflation. In addition, to further differentiate between the conflation hypothesis and the weighted-ΔP hypothesis, we manipulated actual sample size, the total number of trials in a sample.

To control for the influence of outcome density, $P(e)$, which increases as the generative virtual sample size decreases, we manipulated causal direction: Recall that for any particular sample of data, the two causal directions imply complementary virtual sample sizes. Causal direction was manipulated by explicitly and repeatedly instructing participants, between subjects, to consider only one of the two possible causal directions throughout the experiment and for all ratings. Except for causal direction, all variables were manipulated within subjects.

Differentiating Predictions

Recall that according to weighted ΔP, causal-strength ratings reflect weights on the conditional probabilities and are not the result of a conflation of strength and confidence. Therefore, although weighted ΔP is often proportional to virtual sample size, it does not predict any influence of actual sample size on estimates of strength, nor does it predict any correlation between (a) participants' initial-confidence ratings and (b) the deviations of their strength estimates from causal-power predictions.

In contrast, the conflation hypothesis predicts that causal-strength ratings will vary with factors that influence confidence, in particular, both virtual and actual sample size. It also predicts that, conversely, final-confidence ratings will vary with causal-strength estimates. The conflation implies that although causal-strength ratings may deviate from causal-power predictions, these deviations will be correlated with the initial-confidence ratings. Moreover, if our measure of perceived causal strength is sufficiently unambiguous, it should allow some participants to reveal their discrimination between causal strength and confidence due to sample size. Note that the conflation of strength and confidence by some participants may occur side-by-side with dissociation between these concepts in other participants. If so, mean judgments will show evidence for both conflation and dissociation.

Method

Participants

Fifty-four undergraduates at the University of California, Los Angeles, participated to obtain course credit in an introductory psychology course. Two participants who failed to complete all of the conditions in the experiment were excluded from our analyses.

Design

The direction of the causal influence was the only variable manipulated between subjects. Within subjects, two actual sample sizes, 16 and 64, were combined with two causal powers, .25 and 1, and three base rates, that is, $P(e|{\sim}c)$: 0 or 1 (for the generative and preventive condition, respectively), .25, and .75. To isolate the influence of virtual sample size from that of actual sample size, we operationally defined three values of virtual sample size (small, medium, and large) within an actual sample size, as a function of the base rate of *e* and causal direction: Within a particular actual sample, an increase in the base rate of *e* implies a decrease in virtual sample size for a generative cause and a concomitant increase in virtual sample size for a preventive cause. Thus, base rates of .75, .25, and 0 (in decreasing order) in the generative condition correspond respectively to small, medium, and large virtual sample sizes; conversely, base rates of .25, .75, and 1.0 (in increasing order) in the preventive condition correspond respectively to small, medium, and large virtual sample sizes. For the generative and preventive conditions, respectively, the 0 and 1 levels of the base rate variable can be thought of as optimal (having the largest possible virtual sample size for the sample). The .25 and .75 levels play complementary roles for the two causal directions, as should be clear.

Note that virtual sample size, as we operationally define it, is on an ordinal scale. Therefore, statistical interactions involving this variable can be artifacts. For example, when the actual sample size is 64, the virtual sample size increases from 16 (small) to 64 (large), whereas when the actual sample size is 16, the virtual sample size increases from 4 (small) to 16 (large). Consequently, on the basis of the greater difference between virtual-sample frequencies when actual sample size is 64 (a difference of 48) than when it is 16 (a difference of 12), one would expect the influence of virtual sample size to be greater when the actual sample size is 64 than when it is 16.[6]

There were three dependent measures: (a) initial confidence based on the before condition, (b) causal-strength rating, and (c) final confidence based on both the before and the after conditions.

Materials and Procedure

The stimuli were presented on a computer and booklets were provided for writing causal strength and confidence ratings. The various within-subject conditions were presented in random order for each participant. Throughout the materials, only the terms *produce* and *relieve* varied across the generative and preventive conditions. The presence of the minerals (the candidate causes) was verbally labeled, as well as indicated graphically by differ-

[6] Also note that when considered on a ratio scale, virtual sample size always increases with actual sample size, and thus an influence of the latter variable can potentially be attributed to the former. This apparent confounding is not a problem for any of our goals in the present article. It is not a problem for our goal of showing an influence of virtual sample size on confidence, because the attribution would only provide further support for that influence. Neither is it a problem for our goal of manipulating actual sample size to differentiate between the weighted ΔP hypothesis and the conflation hypothesis, because unlike virtual sample size, weighted ΔP does not change across the two levels of actual sample size. We refer to the total number of trials in a sample as "actual sample size" to distinguish between the virtual sample and a conventional definition of sample size, but none of our goals are dependent on ruling out virtual sample size as an explanation for the apparent influence of actual sample size.

ently colored and patterned backgrounds (see Figure 1). The presence or absence of headache (the effect) was indicated by a frowning or a smiling face, respectively, as in Figure 1.

Cover story and initial-confidence question. First, participants were presented with the following cover story:

> A pharmaceutical company is investigating if an allergy medicine might produce headache as a side effect. The company has hired a team of scientists to conduct an experiment testing all of the minerals that compose the allergy medicine.
>
> Each mineral is tested in a different lab, so the number of patients that have a headache before receiving any mineral, as well as the total number of patients, will vary across groups.
>
> You will see each group of patients, both before and after receiving the particular mineral, and it is your job to evaluate its influence on headache.

After the cover story, participants in the generative condition were given the following general instructions on the judgments to be made:

> Before a patient group receives any minerals, we want you to consider how useful the group is for determining if something *produces* headache, and how confident you can feel about the results from that particular lab.
>
> After a patient group receives their mineral, you will be asked about its influence on headache and, again, how confident you feel about the results from that particular lab.

Participants were then presented with 12 randomly ordered consecutive trials, each of which made up of two separate screens illustrating the same patient group.

On the first screen of each trial was a panel representing a particular patient group before they had received any minerals (e.g., top panel in Figure 1A), and participants were asked the following two questions regarding this before condition: "Given the size and characteristics of this particular patient group, how useful do you think it is for determining if the mineral *produces* headache?" and "Given your answer to Question 1, how confident will you feel about the results from this lab?"

Each question was followed by a scale, ranging from 0 to 100, on which participants rated usefulness and confidence, respectively. On the usefulness scale, the 0 point was labeled *not at all useful* and the 100 point was labeled *extremely useful*, whereas on the confidence scale, the 0 point was labeled *not at all confident* and the 100 point was labeled *extremely confident*. The usefulness rating was not analyzed but was included to (a) prompt the relevant causal direction and (b) reduce confusion that might arise from being asked to rate confidence about the results before any results had been displayed.

Strength and final-confidence questions. On the second screen of each trial were two panels. The top panel showed the just viewed before condition while the bottom panel showed the same patient group as they were after receiving the particular mineral (the after condition, e.g., bottom panel of Figure 1A). While viewing this screen, participants were asked the following causal-strength questions:

> What effect does this mineral have on headache?
>
> □ This mineral has ABSOLUTELY NO INFLUENCE on headache.
>
> □ This mineral PRODUCES headache.
>
> ● Suppose that there are 100 people who do not have headaches.
>
> ● If this mineral was given to these 100 people, how many of them would have a headache? __________.[7]

To enable participants to reveal their ability to differentiate between causal strength and confidence due to sample size, we posed a causal-strength question than was less ambiguous that used by Buehner and Cheng (1997). Specifically, we adapted Buehner et al.'s (2003, Experiment 2) causal-strength question. They argued that, compared with rating a subjective perception of an internal state (e.g., "How strongly do you think that *x* causes *y*?"), estimating the frequency of an observable event (e.g., "How many patients would have a headache?") is less likely to be confused with reliability (Buehner et al., 2003). Participants were instructed not to assign a numerical rating when checking the "no influence" box, as well as to not assign a rating of 0 if they checked the "produces" box.

The causal-strength question was followed by a second confidence question that was directly analogous to the initial-confidence question: "Now that you've seen all the information, how confident do you feel about the results from this lab?" The same confidence scale was used.

Results and Discussion

Overview

All three mean ratings (initial confidence, strength, and final confidence) increased with an increase in virtual as well as actual sample sizes. In addition, the strength and final confidence ratings varied with causal power. Because of the influence of causal power on the final confidence rating, the two confidence ratings differed significantly. Finally, an asymmetry between the two causal directions was observed for all three mean ratings: a change in virtual sample size influenced mean ratings more in the generative condition than in the preventive condition.

The rest of this section is organized as follows: In each of three subsections, we report the results from analyses of variance performed on each of our three measures (initial confidence, causal strength, and final confidence). We also address the main issue of our article, the relation between confidence and causal strength, evaluating whether (a) deviations from causal power observed in causal-strength ratings can be explained by the initial-confidence ratings and (b) whether causal strength and confidence are dissociable. We discuss the results in terms of the two hypotheses under consideration: the weighted-ΔP hypothesis and the conflation hypothesis.

All ratings from the generative and preventive conditions were coded as positive numbers, as they were recorded by the participants (without changing signs for the preventive conditions). The means for all measures and all conditions in Experiment 1 are

[7] In the preventive condition, participants are asked to imagine 100 people who do have headaches and how many would no longer have a headache if given the mineral.

listed in Table 1, together with the values of all variables and the predictions of the models.[8]

The Initial-Confidence Rating

A Causal Direction (2) × Virtual Sample Size (3) × Actual Sample Size (2) × Causal Power (2) analysis of variance on the initial-confidence ratings was performed with all factors except causal direction as within-subject variables. Note that causal power is a dummy variable in this analysis; the before conditions were identical across the two levels of causal power.

Virtual and actual sample size. Figure 2 depicts the mean initial-confidence ratings for all relevant conditions.[9] The mean ratings clearly increased with virtual sample size, $F(2, 100) = 117.1$, $MSE = 512.2$, $p < .001$, $\eta_p^2 = .70$. Recall that for these ratings, because the after condition has not yet been presented to the participants, ΔP can be excluded as an explanation for this observed trend. Planned comparisons reveal that the mean rating for the conditions with a large virtual-sample size ($M = 73.4$) was significantly higher than that for a medium virtual-sample size ($M = 57.8$), which in turn was significantly higher than that for a small virtual-sample size ($M = 39.5$), MSEs = 1,978.4 and 3,881.6, for the respective comparisons, both Fs > 72.0, both ps < .001, both η_p^2s > .59. The difference between the medium and small virtual-sample conditions is especially important for isolating the effect of virtual sample size: These conditions had equal sampling variability, $SD_{P(e|\sim c)} = \sqrt{P(e|\sim c) \cdot (1 - P(e|\sim c))}$, because their values of the base rate of e are symmetrical around .5.

Figure 2 also shows an influence of actual sample size: The mean initial-confidence rating was higher when the actual sample size was large ($M = 64.3$) than when it was small ($M = 49.5$), $F(1, 50) = 63.56$, $MSE = 540.5$, $p < .001$, $\eta^2 = .56$. This confirms that our initial-confidence rating is indeed sensitive to sample size. In addition, as expected due to our operational definition of virtual sample size, there was an interaction between actual and virtual sample size (see the *Design* section), $F(2, 100) = 5.7$, $MSE = 226.4$, $p < .005$, $\eta_p^2 = .10$, such that the difference between the mean initial-confidence ratings for a small and large virtual sample size was greater when the actual sample size was 64 (mean difference = 38.7) than when the actual sample size was 16 (mean difference = 29.2); mean square errors for the differences were 2,220.9 and 1,753.8, respectively, Fs > 101.0, ps < .001, η_p^2s > .66. As discussed earlier, this interaction is artifactual.

Causal direction: An asymmetry in the effect of virtual sample size. There appears to be an asymmetry between the two causal directions such that the difference between mean initial confidence ratings for small and large virtual sample sizes was greater in the generative condition (mean difference = 39.9) than in the preventive condition (mean difference = 28.0); mean square errors for the differences were 7,594.0 and 5,273.0, respectively, Fs > 61.0, ps < .001, η_p^2s > .71, as reflected in an interaction between virtual sample size and causal direction, $F(2, 100) = 3.7$, $MSE = 512.2$, $p < .05$, $\eta_p^2 = .07$. This asymmetry may be interpreted in terms of the presence of alternative causes: Whereas a high base rate of e raises doubts about attributing e to c in view of the presence of alternative generative causes, a low base rate of e does not lead to any corresponding doubts due to alternative preventive causes. Increasing this base rate therefore reduces confidence more for generative causes than it increases confidence for preventive causes.[10]

Ratio of virtual to actual sample size. So far, the influence of virtual sample size has been assessed on an ordinal scale within a particular actual sample size. However, when this influence was assessed on a ratio scale across different actual sample sizes, there was an unexpected influence of the ratio of virtual sample size to actual sample size, which we term R. Consider two before conditions across which actual and virtual sample size both decrease (from 64 to 16 and from 16 to 12, respectively), but R increases from .25 to .75. For the generative case, in spite of the decrease in both actual and virtual sample sizes, the mean initial-confidence ratings increased significantly across these two conditions, from 41.5 to 50.6, $F(1, 51) = 5.6$, $MSE = 382.0$, $p < .05$, $\eta_p^2 = .10$, whereas for the two analogous preventive conditions, mean initial-confidence ratings did not differ significantly, $F(1, 51) = 1.4$, $MSE = 242.6$, $p = .25$, $\eta_p^2 = .03$.

Also note that ΔP, which is directly proportional to R, fails to explain any variations in the initial confidence ratings, as well as the asymmetry between causal directions. Instead, as with the interaction between virtual sample size and causal direction, we interpret these results in terms of doubts about attributing e to c in view of the presence of alternative generative causes. The complement of R (i.e., the base rate of e) may be regarded as an estimate of the probability at which causes other than the candidate produce e (see Cheng, 1997). Therefore, the higher the R, the more confident one can be that e is not due to other causes.

The causal-strength rating. The "no influence" option was chosen for 78 out of the total 624 ratings, and these judgments were coded 0 (participants did not assign a numerical rating when selecting this option). A Causal Power (2) × Causal Direction (2) × Virtual Sample Size (3) × Actual Sample Size (2) analysis of variance was performed on the causal-strength ratings, with all factors except causal direction as within-subject variables.

All four main effects are significant. Figure 3 depicts the mean causal-strength ratings as a function of virtual sample size for each

[8] We computed weighted-ΔP predictions using the same weights on conditional probabilities as those used by Lober and Shanks (2000); 1.0 to $P(e|c)$ and 0.6 to $P(e|\sim c)$. It should be noted that with other weights—for example, 1.0 and 0.8 for $P(e|c)$ and $P(e|\sim c)$, respectively—the weighted-ΔP values become more proportional to virtual sample size.

[9] Recall that the size of the interval between the small and medium virtual sample size is greater than that between a medium and large virtual sample size. This is not indicated on the x-axis of Figure 2 or any subsequent figure.

[10] This interpretation is loosely consistent with the "no preventive background cause" assumption in the power-PC theory: Whereas the theory explains a positive base rate of e by e being produced by generative background causes, it does not explain a base rate of e less than the maximum (i.e., less than 1) by preventive background causes. That is, the presence of e requires an explanation, but the absence of e requires no explanation. Note, however, that this theory does not in itself predict a difference in causal strength between the analogous conditions for the two causal directions. The loose interpretation would require an extension of the causal approach underlying the theory to account for confidence in causal-strength estimates or in judgments regarding the existence of a causal relation (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, in press).

Table 1
Means and Standard Deviations of the Three Dependent Measures for All Conditions in Experiment 1, Together With The Values of Independent Variables

				Initial conf		Strength		Final conf			
Pwr	Ss	Vss	Dir	*M*	*SD*	*M*	*SD*	*M*	*SD*	ΔP	WΔP
.25	16	4	G	29.69	22.92	12.23	20.55	35.42	25.5	0.06	0.36
.25	16	4	P	35.19	25.86	13.58	17.95	32.31	23.97	−0.06	0.04
.25	16	12	G	48.81	19.37	19.12	13.46	42.54	21.12	0.19	0.29
.25	16	12	P	50.39	20.34	16.73	14.13	39.62	23.32	−0.19	0.11
.25	16	16	G	65.58	24.08	20.15	19.4	52.5	22.57	0.25	0.25
.25	16	16	P	58.85	24.71	15.81	10.5	46.81	21.47	−0.25	0.15
.25	64	16	G	43.69	24.23	8.39	10.04	41.27	18.82	0.06	0.36
.25	64	16	P	48.85	24.79	23.96	25.4	46.35	23.05	−0.06	0.04
.25	64	48	G	61.35	16.34	20.73	12.07	50.35	18.15	0.19	0.29
.25	64	48	P	66.35	13.75	27.15	22.01	57.5	19.71	−0.19	0.11
.25	64	64	G	85.58	9.2	21	8.03	68.46	19.48	0.25	0.25
.25	64	64	P	81.92	18.06	30.58	19.97	63.08	23.11	−0.25	0.15
1	16	4	G	33.35	24.61	46.04	41.8	42.15	27.27	0.25	0.55
1	16	4	P	37.69	29.71	65.5	42.88	45.58	30.41	−0.25	−0.15
1	16	12	G	52.39	20.6	80.96	19.6	66.08	19.01	0.75	0.85
1	16	12	P	53.73	18.96	87.5	20.7	63.46	23.48	−0.75	−0.45
1	16	16	G	69.35	22.85	90.96	20.05	81.08	12.67	1	1
1	16	16	P	58.92	21.81	93.27	17.14	69.23	24.5	−1	−0.6
1	64	16	G	39.39	22.34	53.31	39.71	47.62	25.41	0.25	0.55
1	64	16	P	48.08	22.05	79.62	33.61	62.89	21.5	−0.25	−0.15
1	64	48	G	64.89	15.14	83.65	17.86	79.65	9.73	0.75	0.85
1	64	48	P	64.81	18.79	92.5	11.09	84.81	12.53	−0.75	−0.45
1	64	64	G	85	12.81	98.65	4.37	86.81	14.18	1	1
1	64	64	P	82.12	16.01	98.46	3.09	90.65	11.9	−1	−0.6

Note. Pwr = causal power; Ss = actual sample size; Vss = virtual sample size; Dir = causal direction; conf. = confidence; G = generative; P = preventive. The last two columns list the predictions of ΔP and weighted ΔP (WΔP).

causal power, with causal direction and actual sample size as parameters. As predicted by both hypotheses under consideration, these ratings varied with virtual sample size (*M*s = 37.8, 53.5, and 58.6 for small, medium, and large virtual sample sizes, respectively), *F*(2, 100) = 44.3, *MSE* = 551.0, $p < .001$, $\eta_p^2 = .47$. Planned comparisons reveal that both pairwise differences between adjacent virtual-sample-size conditions are significant, *MSE*s = 1,704.4 and 4,738.3, respectively, *F*s > 12.0, *p*s < .005, η_p^2s > .2. As predicted by the conflation hypothesis but not the weighted-ΔP hypothesis, strength ratings also increased with an increase in actual sample size (*M*s = 46.8 and 53.2 for actual sample sizes of 16 and 64, respectively), *F*(1, 50) = 16.5, *MSE* = 380.5, $p < .001$, $\eta_p^2 = .25$. Strength ratings also increased with an increase in power (*M*s = 19.1 and 80.9 for

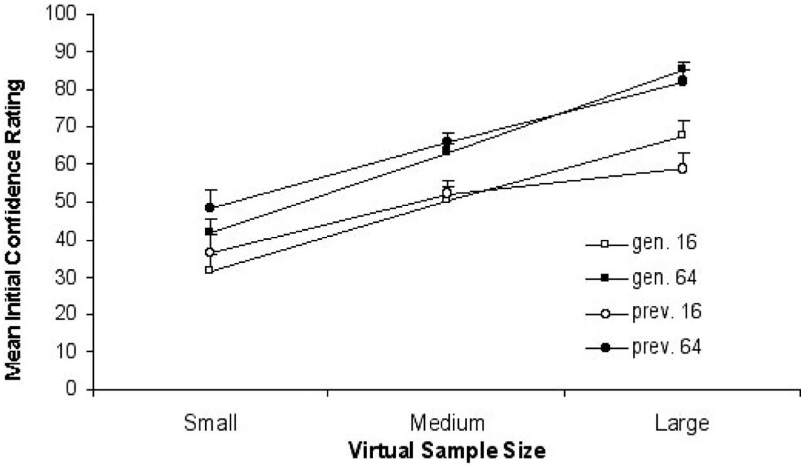

Figure 2. Experiment 1: Mean initial confidence ratings as a function of virtual sample size with causal direction and actual sample size as parameters. Error bars represent standard error of the mean. gen. = generative; prev. = preventive.

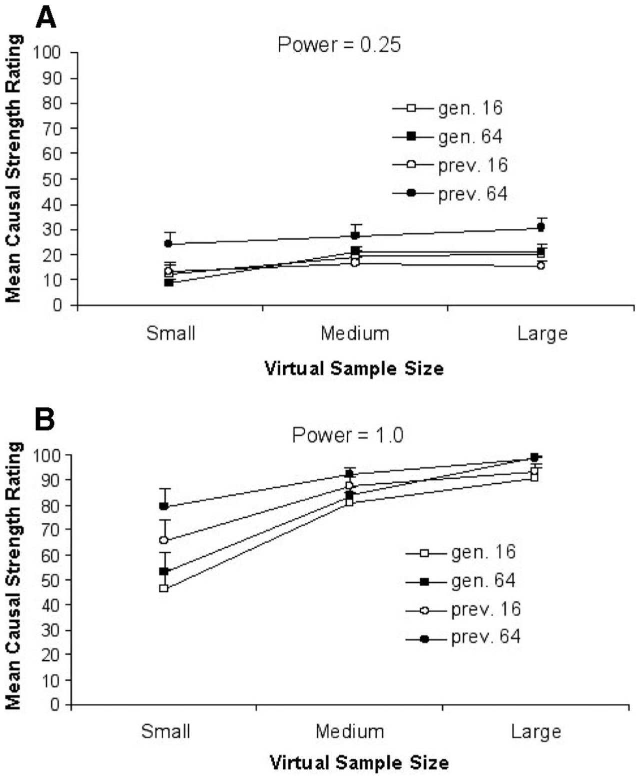

Figure 3. Experiment 1: Mean causal-strength ratings as a function of virtual sample size with causal direction and actual sample size as parameters, for powers of (A) .25 and (B) 1.0. Error bars represent standard error of the mean. gen. = generative; prev. = preventive.

causal power of .25 and 1.0, respectively), $F(1, 50) = 915.6$, $MSE = 649.7$, $p < .001$, $\eta_p^2 = .95$. Finally, mean causal-strength ratings varied with causal direction (Ms = 46.3 and 53.7 in the generative and preventive conditions, respectively), resulting in a main effect of causal direction, $F(1, 50) = 5.58$, $MSE = 1{,}553.0$, $p < .05$, $\eta_p^2 = .10$.

Neither the weighted ΔP model nor the conflation hypothesis predicts an influence of causal direction. Only the conflation hypothesis can explain the influence of actual sample size, but both hypotheses explain the other two main effects.

Interactions. There were three reliable interactions. As with the initial-confidence rating, there was an interaction between virtual sample size and causal direction, $F(2, 100) = 5.0$, $MSE = 551.0$, $p < .01$, $\eta_p^2 = .09$; the difference between mean causal-strength ratings for small and large virtual sample sizes was greater in the generative condition (mean difference = 27.7) than in the preventive condition (mean difference = 12.9), MSEs = 6,845.6 and 6,716.2, respectively, Fs > 11.0, ps < .005, η_p^2s > .32. In contrast, the difference between causal-strength ratings due to actual sample size was smaller in the generative condition than in the preventive condition, resulting in an interaction between actual sample size and direction, $F(1, 50) = 6.8$, $MSE = 380.5$, $p < .05$, $\eta_p^2 = .09$; this interaction, however, was not replicated in a follow-up experiment (see Experiment 2) and is not discussed further.

Finally, there was an interaction between virtual sample size and power, $F(2, 100) = 19.3$, $MSE = 510.4$, $p < .001$, $\eta_p^2 = .28$; the difference between causal-strength ratings for small and large virtual sample sizes was greater when power was 1.0 (mean difference = 34.2) than when power was .25 (mean difference = 7.4), MSEs = 5,179.9 and 1,287.7, respectively, Fs > 8.0, ps < .01, η_p^2s > .14. This interaction appears to be a floor effect. Notice that the mean difference when power was 1.0, which was nearly 35 points, was greater than the maximum possible (reasonable) difference when power was .25; at the lower power value, both major models, ΔP and causal power, predict a causal-strength rating no higher than 25. The analogous maximum possible difference when power is 1.0 is 100 points.

The relationship between initial confidence and deviations from strength predictions. Recall that a goal of measuring causal strength was to test whether the observed deviations from causal power predictions can be explained by the initial-confidence ratings. To further evaluate the conflation hypothesis, we assessed whether the initial-confidence ratings accounted for the deviation of causal-strength ratings from causal-power predictions. For each of the 24 conditions, we obtained the difference between the mean causal-strength rating and the power prediction. We performed a linear regression of these differences on the mean initial confidence ratings and found that r was quite high, both for conditions with a causal power of .25 ($r = .72$), $F(1, 10) = 10.9$, $MSE = 20.9$, $p < .01$, and for those with a causal power of 1.0 ($r = .89$), $F(1, 10) = 36.1$, $MSE = 71.2$, $p < .001$. These relations are inexplicable by ΔP (weighted or unweighted), because the initial-confidence ratings were measured before the contingency necessary for calculating ΔP was available. They strongly support the conflation hypothesis: Variations in confidence due to variations in virtual sample size led to the deviations from causal power observed in causal-strength ratings. (Further support for our conclusion comes from an evaluation of Bayesian models of causal-structure selection that corresponds to ΔP and causal power, the two models of strength; Lu et al., in press.)

It would be misleading, however, to base conclusions solely on the condition means. There was clear evidence for two modes in the strength-rating distributions of conditions in which causal power was 1.0 and ΔP was .25 or .75. For these conditions, a small proportion of participants gave ratings that coincided with ΔP predictions; specifically, out of 208 such ratings, 27 ratings were exactly as predicted by ΔP. For comparison, 99 ratings were exactly as predicted by power. For conditions in which power was .25, no bimodality was observed. This may be due to the fact that for those conditions, (a) the differences between causal power and ΔP predictions were smaller and (b) many subjects chose the "no influence" option.

Recall that, when causal power is 1.0, ΔP values of .25 and .75 correspond to small and medium virtual sample sizes, respectively. The difference in mean strength ratings between conditions with a ΔP of .25 (i.e., small virtual sample size; $M = 86.2$) and those with a ΔP of .75 (i.e., medium virtual sample size; $M = 61.1$) was 25.0, $F(1, 50) = 34.0$, $MSE = 959.2$, $p < .001$, $\eta_p^2 = .41$. Note that, contrary to the conflation hypothesis, the bimodality clearly indicates that this difference was at least in part due to participants rating according to ΔP rather than to the influence of virtual sample size on confidence. However, this difference remains substantial even when those 27 ratings that are numerically predicted by ΔP are excluded from the analysis; the mean difference between conditions with small and medium virtual sample sizes for the remaining 181 ratings was 21.6, which is 86% of the total mean difference of 25.0. To evaluate the difference due to virtual sample size statistically, we excluded all participants who gave at least one rating that equaled ΔP; the mean difference between the relevant conditions for the 36 remaining participants (144 remaining ratings) was 16.0, $F(1, 34) = 10.0$, $MSE = 882.6$, $p < .005$, $\eta_p^2 = .23$. In other words, the bimodality of the distribution may exaggerate but does not fully account for the evident influence of virtual sample size on causal-strength estimates.

The final-confidence rating. As for causal-strength ratings, a Causal Direction (2) × Virtual Sample Size (3) × Causal Power (2) × Actual Sample Size (2) analysis of variance was performed on the final-confidence ratings with all factors except causal direction as within-subject variables. Except for causal direction, all main effects were significant. Recall that although neither ΔP nor causal power explains confidence ratings, these three main effects are consistent with the conflation hypothesis. It is unclear how the weighted ΔP model can explain them. Figure 4 depicts the mean final-confidence ratings as a function of virtual sample size for each causal power, with causal direction and actual sample size as parameters. There was a clear influence of virtual sample size on the final-confidence ratings (*M*s = 69.8, 60.5, and 44.2 for large, medium, and small virtual sample sizes, respectively), $F(2, 100) = 91.8$, $MSE = 381.2$, $p < .001$, $\eta_p^2 = .65$. A planned comparison revealed that each pairwise difference between adjacent levels was significant, *MSE*s = 2,694.7 and 1,758.3 for the respective comparisons, $F(1, 50) > 41.0$, $p < .001$, $\eta_p^2 = .45$. In addition, there was a clear influence of actual sample size (*M*s = 51.4 and 65.0 for actual sample sizes of 16 and 64, respectively), $F(1, 50) = 62.9$, $MSE = 456.1$, $p < .001$, $\eta_p^2 = .56$.

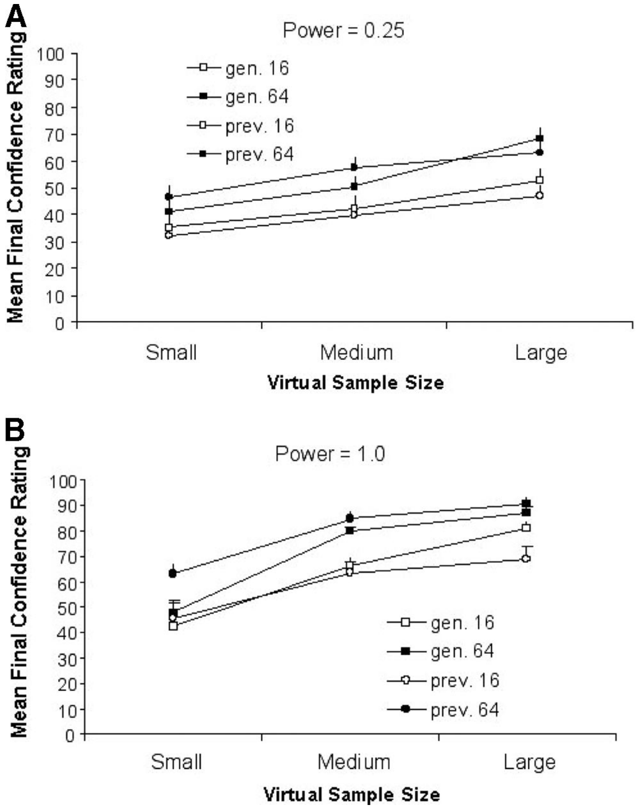

Figure 4. Experiment 1: Mean final confidence ratings as a function of virtual sample size with causal direction and actual sample size as parameters, for a power of (A) .25 and (B) 1.0. Error bars represent standard error of the mean. gen. = generative; prev. = preventive.

Causal power also influenced the final-confidence ratings (*M*s = 48.0 and 68.3 for causal powers of .25 and 1.0, respectively), $F(1, 50) = 102.6$, $MSE = 627.5$, $p < .001$, $\eta_p^2 = .67$. This influence might reflect a conflation of perceived strength and confidence or a direct influence of the sampling variability of strength on confidence. A causal power of 1.0 has less sampling variability than a causal power of .25.

Interactions. As with the initial-confidence and strength ratings, the final-confidence ratings showed the asymmetry between causal directions: The difference between mean final-confidence ratings for the small and large virtual sample sizes was greater in the generative condition (mean difference = 30.6) than in the preventive one (mean difference = 20.7), *MSE*s = 6,060.6 and 3,331.0 for the respective comparisons, *F*s > 53.0, *p*s < .001, η_p^2s > .68, resulting in an interaction between virtual sample size and direction, $F(2, 100) = 3.5$, $MSE = 381.2$, $p < .05$, $\eta_p^2 = .07$.

There were two additional interactions, neither of which is theoretically interesting. First, the difference between final-confidence ratings due to actual sample size was smaller in the generative condition than in the preventive one, resulting in an interaction between actual sample size and direction, $F(1, 50) = 6.9$, $MSE = 456.1$, $p < .05$, $\eta_p^2 = .12$. As with the similar interaction observed for causal-strength ratings, this interaction was not replicated and hence is not discussed.

Second, there was a Virtual Sample Size × Causal Power interaction, $F(2, 100) = 16.5$, $MSE = 220.2$, $p < .001$, $\eta_p^2 = .25$. The difference between mean final-confidence ratings for the small and large virtual sample sizes was greater when power equaled 1.0 (mean difference = 32.4) than when power equaled .25 (mean difference = 18.9), *MSE*s = 1,780.9 and 1,619.3, respectively, *F*s > 45.0, *ps* < .001, η_p^2s > .48. This interaction is likely due to the floor effect observed for causal-strength ratings, discussed previously: When power equaled .25 and the virtual sample size was small, a substantial proportion of the strength ratings were 0 (40 out of 104); that is, participants selected the "no influence" option. It appears that the binary decision of checking the "no influence" box, relative to picking one specific number from a continuous range, led to higher confidence, thereby narrowing the difference between the small and large virtual sample size conditions. This might have produced what is essentially a floor effect even though the mean final-confidence ratings were not near 0.

A comparison with initial-confidence ratings. To directly assess the relationship between the initial and final confidence ratings, a Causal Power (2) × Causal Direction (2) × Virtual Sample Size (3) × Actual Sample Size (2) × Question (2) analysis of variance was performed with all factors except causal direction as within-subject variables.

This analysis yielded no main effect of question. Moreover, none of the interactions were theoretically interesting. There were only the expected interactions involving the question variable due to causal power being an irrelevant dummy variable for the initial-confidence ratings. Accordingly, there was an interaction between power and question: The final, but not the initial, confidence rating varied with power, $F(1, 50) = 84.2$, $MSE = 341.4$, $p < .001$, $\eta_p^2 = .63$. Likewise, whereas the influence of virtual sample size on the final confidence rating was attenuated when power was low, no such attenuation was observed for the initial confidence rating, resulting in a two-way interaction between virtual sample size and question, $F(2, 100) = 10.4$, $MSE = 187.3$, $p < .001$, $\eta_p^2 = .17$, and a three-way interaction between virtual sample size, power, and question, $F(2, 100) = 10.1$, $MSE = 147.2$, $p < .001$, $\eta_p^2 = .17$.

Finally, recall that there was an (unreplicated) interaction between actual sample size and causal direction for the final-confidence rating. Because there was no such interaction for the initial-confidence rating, the current analysis revealed a three-way interaction between actual sample size, causal direction, and question, $F(1, 50) = 4.5$, $MSE = 169.2$, $p < .05$, $\eta_p^2 = .08$. No other interactions involving the question variable were significant.

The ratio of virtual to actual sample size. Recall the two sets of conditions across which both virtual and actual sample sizes increased but *R*, the ratio of virtual to actual sample size, decreased. The pattern of results found for the initial confidence ratings held true for the other two ratings in Experiment 1: Collapsed across the causal power variable (to avoid the previously discussed floor effect), both mean causal-strength ratings and final-confidence ratings increased with *R* across the two generative conditions, *MSE*s = 384.2 and 377.4, respectively, *F*s > 6.5, *p*s < .02, η_p^2s > .21, but did not differ significantly for the two analogous preventive conditions, *MSE*s = 412.0 and 166.7, respectively, *F*s < 1.5, *p*s > .23, η_p^2s < .06.

Dissociability of strength and confidence. Our results show clear evidence for the dissociability of strength estimates from confidence in those estimates. First, the modal causal-strength ratings were exactly as predicted by causal power for all except two conditions in which the virtual and actual sample sizes were both small and causal power was low. Even for these two conditions, in which the modal answer was “no influence,” causal power was the modal numerical rating when a numerical rating was given. Second, across our two most extreme sample size conditions (i.e., actual and virtual sample sizes were both small vs. both large) for which causal power was 1, more than 40% of the participants (21 out of 50) rated strength at 100 for both conditions. Note that there is a 16-fold increase in virtual sample size across these conditions. The influence of sample size for this subset of participants was expressed exclusively in their confidence ratings: Across these conditions, their mean final-confidence ratings increased enormously from 49.1 to 90.7, $F(1, 20) = 34.8$, $MSE = 524.2$, $p < .001$, $\eta_p^2 = .64$. Moreover, even those participants whose strength estimates did deviate from causal power in the direction predicted by sample size showed evidence for dissociation; their strength ratings differed significantly from their final confidence ratings in both of the two extreme conditions, *MSE*s = 600.1 and 84.8 for the respective comparisons, *F*s > 4.5, *p*s < .05, η_p^2s > .14.

Third, in view of the less ambiguous wording of our causal-strength question, participants’ ratings of strength and confidence varied in opposite directions when causal power and virtual sample size exerted pressures in opposite directions. Consider two extreme conditions in our experiment in which (a) power was .25 and virtual-sample size was 64 and (b) power was 1.0 and virtual-sample size was 4. Because ΔP remained constant at .25 and sampling variability was equated, these variables are eliminated as explanations for any differences observed across the two types of conditions. Figure 5 shows the observed means for all three measures for these two conditions, collapsed across causal direction. A Condition (2) × Measure (3) analysis of variance revealed that the apparent interaction between condition and measure was highly significant, $F(2, 102) = 104.3$, $MSE = 415.6$, $p < .001$, $\eta_p^2 = .67$. Specifically, whereas both the initial and final-confidence ratings decreased, *MSE*s = 328.7 and 626.8, respectively, *F*s > 4.00, *p*s < .05, η_p^2s > .08, as would be expected with the decrease in virtual sample size, the causal-strength ratings increased, $F(1, 51) = 52.2$, $MSE = 823.6$, $p < .001$, $\eta_p^2 = .51$, as would be expected with the increase in power.

Summary and Conclusions

In summary, the initial-confidence ratings clearly varied with virtual sample size and accounted quite well for the deviations from causal power predictions observed in strength ratings, indicating that there was a conflation of strength and confidence. At the same time, there were several lines of converging evidence for a dissociation of strength estimates and confidence in those estimates. For example, across conditions at two extremes in which causal power and sample size varied in opposite directions, observed causal-strength and confidence ratings likewise changed in opposite directions.

Our results clearly favor the conflation hypothesis over the weighted-ΔP hypothesis. Only the former can explain (a) the influence of actual sample size on strength ratings, (b) the correlation between the initial-confidence ratings and the deviations of causal-strength ratings from causal power predictions, and (c) the primary mode coinciding with causal power for strength ratings.

However, some aspects of the results still pose a challenge for the conflation hypothesis. Most notably, this hypothesis cannot explain why a small subset of participants rated strength according to ΔP. Nor can it explain why, despite our less ambiguous causal-strength question, a large number of participants still apparently conflated strength and confidence. We return to these aspects of our results in the General Discussion section. But first, in Experiment 2, we assess whether the initial confidence rating had any influence on causal-strength ratings and the final confidence ratings.

Experiment 2: A Replication Without the Initial-Confidence Question

It may be argued that we encouraged participants to consider virtual sample size by asking them to evaluate the usefulness of a sample given its size and characteristics and their confidence in the results. These questions may also seem rather odd, coming as they did before data from the after condition had been revealed. In addition, the final-confidence question, which was designed to closely match the initial-confidence question (it therefore asked about confidence in “results from this lab”), may have seemed too ambiguous to be meaningfully assessed. To evaluate the impact of these potential problems, we conducted a follow-up experiment that omitted the initial-confidence question and used a more clearly worded confidence measure.

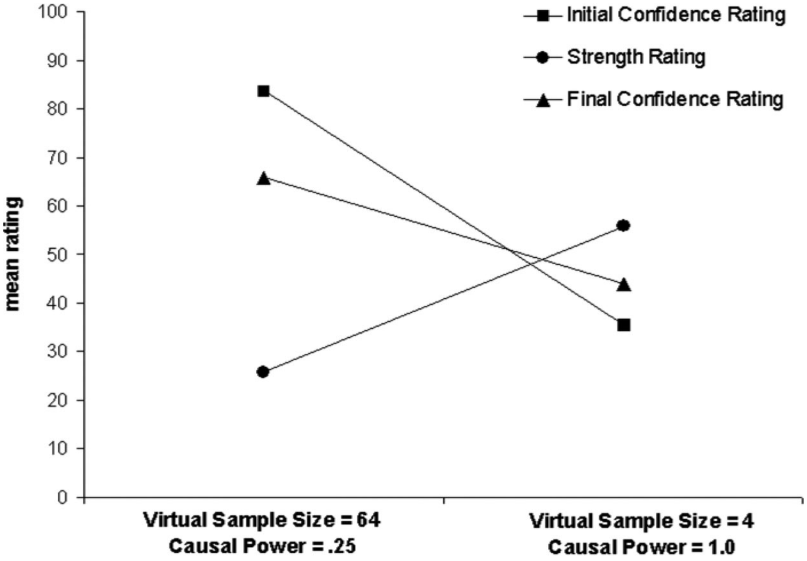

Figure 5. Experiment 1: Mean ratings of all three measures for conditions across which virtual sample size and ΔP decrease while power increases, collapsed over actual sample size and causal direction.

Method

Participants

Forty University of California, Los Angeles, undergraduates participated in the replication experiment to obtain credit in an introductory psychology course. Two participants failed to complete all of the conditions and were excluded from our analysis.

Materials and Procedure

The materials and procedures of the replication experiment were identical to those of Experiment 1 except that the usefulness question and initial-confidence question, as well as any instructions pertaining to these two measures, were omitted. Participants were presented with the before and after conditions simultaneously and were simply asked to first estimate causal strength using the same strength question as in the main experiment. They were then asked to rate how confident they were in their response to the strength question (i.e., “How confident are you in response to Question 1?”) on the same confidence scale as was used in Experiment 1.

Results and Discussion

The means for all measures and all conditions in Experiment 2 are listed in Table 2, together with the values of all variables and the predictions of the models. We performed the same analyses on the causal-strength and confidence ratings as we did in Experiment 1. Notably, every finding in our main experiment relevant to our conflation and dissociation hypotheses was replicated, as detailed below. All conclusions in the main experiment are therefore supported by our replication results.

The only results that did not replicate were interactions that appear to be artifacts and extraneous to our hypotheses; these failures were the ones mentioned without discussion earlier. In addition, there was an interaction that did not appear in our previous experiment, as we report below.

Strength Ratings

The modal numerical strength ratings for all conditions were as predicted by causal power. All main effects except causal direction, $F(1, 36) = 2.5$, $MSE = 679.5$, $p = .12$, $\eta_p^2 = .07$, were significant.

Figure 6 depicts the mean strength ratings as a function of virtual sample size, with causal direction and actual sample size as parameters, separately for each of the two causal powers. Strength ratings increased with causal power (*M*s = 17.5 and 82.2 for causal powers of .25 and 1.0, respectively), $F(1, 36) = 1073.3$, $MSE = 443.6$, $p < .001$, $\eta_p^2 = .97$; actual sample size (*M*s = 46.7 and 53.3 for actual sample sizes of 16 and 64, respectively), $F(1, 36) = 22.8$, $MSE = 242.7$, $p < .001$, $\eta_p^2 = .39$; and virtual sample size (*M*s = 36.7, 53.3, and 59.6 for small, medium, and large virtual sample sizes, respectively), $F(2, 72) = 52.5$, $MSE = 402.7$, $p < .001$, $\eta_p^2 = .59$. Planned comparisons revealed that both pairwise differences between adjacent virtual sample size conditions were significant, *MSE*s = 1,663.2 and 3,466.3 for each comparison, respectively, *F*s > 14.0, *p*s < .005, η_p^2s > .28.

As in Experiment 1, there was an interaction between virtual sample size and power, $F(2, 72) = 19.9$, $MSE = 359.0$, $p < .001$, $\eta_p^2 = .36$. The difference between causal-strength ratings due to virtual sample size was greater when power was 1.0 (*M*s = 61.4 and 97.4 for small and large virtual sample sizes, respectively) than when power was .25 (analogous *M*s = 12.1 and 21.7), *MSE*s = 3,449.4 and 607.3, respectively, *F*s > 23.0, *p*s < .001, η_p^2s > .39. There was also an interaction between virtual sample size and causal direction, $F(2, 72) = 5.9$, $MSE = 402.7$, $p < .005$, $\eta_p^2 = .14$. Specifically, the difference between causal-strength ratings due to virtual sample size was greater in the generative condition (*M*s = 30.3 and 60.5 for small and large virtual sample

Table 2

Means and Standard Deviations of the Two Dependent Measures for All Conditions in Experiment 2, Together With the Values of Independent Variables

				Strength		Confidence			
Pwr	Ss	Vss	Dir	*M*	*SD*	*M*	*SD*	ΔP	WΔP
.25	16	4	G	8.16	10.3	52.63	27.81	0.06	0.36
.25	16	4	P	14.47	11.29	62.89	31.9	−0.06	0.04
.25	16	12	G	14.37	9.21	58.95	24.58	0.19	0.29
.25	16	12	P	15.11	9.43	64.11	31.57	−0.19	0.11
.25	16	16	G	21.84	7.85	65.26	23.66	0.25	0.25
.25	16	16	P	19.05	14.91	62.32	30.44	−0.25	0.15
.25	64	16	G	11.74	10.36	59.21	25.78	0.06	0.36
.25	64	16	P	14	13.91	62.26	32.7	−0.06	0.04
.25	64	48	G	24.68	19.02	61.58	21.8	0.19	0.29
.25	64	48	P	21.11	12.5	60.53	26.65	−0.19	0.11
.25	64	64	G	22.89	9.47	76.05	16.21	0.25	0.25
.25	64	64	P	23.05	11.69	69.47	22.04	−0.25	0.15
1	16	4	G	45.26	40.94	57.11	22.19	0.25	0.55
1	16	4	P	64.47	36.21	70.79	32.41	−0.25	−0.15
1	16	12	G	80.53	23.8	73.68	18.84	0.75	0.85
1	16	12	P	82.11	20.16	75.79	23.76	−0.75	−0.45
1	16	16	G	97.37	7.33	86.68	10.46	1	1
1	16	16	P	93.68	12.89	83.05	17.84	−1	−0.6
1	64	16	G	56.05	34.86	62.89	22.07	0.25	0.55
1	64	16	P	79.74	31.47	74.47	20.06	−0.25	−0.15
1	64	48	G	92.11	12.17	82.47	9.44	0.75	0.85
1	64	48	P	96.05	8.09	84.21	13.46	−0.75	−0.45
1	64	64	G	100	0	95.11	7.26	1	1
1	64	64	P	98.63	4.66	94.42	15.96	−1	−0.6

Note. Pwr = causal power; Ss = actual sample size; Vss = virtual sample size; Dir = causal direction; G = generative; P = preventive. The last two columns list the predictions of ΔP and weighted ΔP (WΔP).

sizes, respectively) than in the preventive condition (analogous *M*s = 43.2 and 56.6), *MSE*s = 4,794.2 and 4,273.9, respectively, *F*s > 16.0, *p*s < .005, η_p^2s > .48.

In contrast to the above replications of our earlier results, the interaction between actual sample size and causal direction was not replicated in the current experiment, *F*(1, 36) = 0.1, *MSE* = 242.7, *p* = .83, η_p^2 = .001. Moreover, there was a significant interaction between actual sample size and power, which did not reach significance in the previous experiment, *F*(1, 36) = 4.2, *MSE* = 227.5, *p* < .05, η_p^2 = .10. Because these interactions were not observed across experiments, they are not discussed further.

Confidence Ratings

Figure 7 depicts the mean confidence ratings as a function of virtual sample size, with causal direction and actual sample size as parameters, separately for each of the causal powers. Confidence ratings increased significantly with causal power (*M*s = 62.9 and 78.4 for causal powers of .25 and 1.0, respectively), *F*(1, 36) = 44.0, *MSE* = 618.9, *p* < .001, η_p^2 = .55; actual sample size (*M*s = 67.8 and 73.6 for actual sample sizes of 16 and 64, respectively), *F*(1, 36) = 10.7, *MSE* = 357.6, *p* < .005, η_p^2 = .23; and virtual sample size (*M*s = 62.8, 70.2, and 79.1 for small, medium, and large virtual sample sizes, respectively), *F*(2, 72) = 52.5, *MSE* = 311.3, *p* < .001, η_p^2 = .47. Planned comparisons again revealed that both pairwise differences between adjacent virtual-sample-size conditions were significant, *MSE*s = 1,570.7 and 1,949.3 for respective comparison, *F*s > 16.0, *ps* < .001, η_p^2s > .32.

There was an interaction between virtual sample size and power, *F*(2, 72) = 6.6, *MSE* = 326.8, *p* < .005, η_p^2 = .15: The difference between confidence ratings due to virtual sample size was greater when power was 1.0 (*M*s = 66.3 and 89.8 for small and large virtual sample sizes, respectively) than when power was .25 (analogous *M*s = 59.3 and 68.3), *MSE*s = 1,737.2 and 1,955.0, respectively, *F*s > 6.0, *p*s < .05, η_p^2s > .14. There was also an interaction between virtual sample size and causal direction, *F*(2, 72) = 5.3, *MSE* = 311.3, *p* < .01, η_p^2 = .13. The difference between confidence ratings due to virtual sample size was greater in the generative condition (*M*s = 58.0 and 80.8 for small and large virtual sample sizes, respectively) than in the preventive condition (analogous *M*s = 67.6 and 77.3), *MSE*s = 5,629.8 and 2,272.5 for respective comparisons, *F*s > 12.0, *p*s < .005, η_p^2s > .41. In contrast to the above replications of our previous results, the interaction between actual sample size and causal direction did not replicate, *F*(1, 36) = 0.6, *MSE* = 357.6, *p* = .44, η_p^2 = .02.

Conflation and Dissociation of Strength and Confidence in Strength

To further assess whether confidence levels can account for the deviation of causal-strength ratings from causal-power predictions, we performed a cross-experiment linear regression of the deviations observed in Experiment 2 on the mean initial confidence ratings from Experiment 1. As in Experiment 1, *r* was quite high, both for conditions with a causal power of .25 (*r* = .88), *F*(1, 10) = 34.4, *MSE* = 6.8, *p* = .001, and for those with a causal

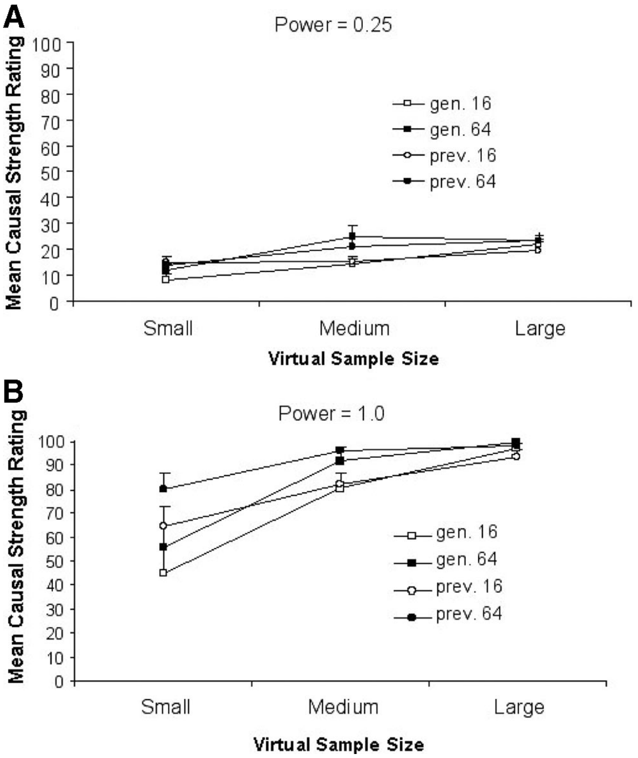

Figure 6. Experiment 2: Mean causal-strength ratings as a function of virtual sample size with causal direction and actual sample size as parameters, for a power of (A) .25 and (B) 1.0. Error bars represent standard error of the mean. gen. = generative; prev. = preventive.

power of 1.0 (r = .91), $F(1, 10) = 47.4$, $MSE = 62.9$, $p = .001$. These results provide further support for the hypothesis that variations in confidence due to variations in virtual sample size explain the deviations from causal power observed in causal-strength ratings.

The ratings from Experiment 2, just as those in Experiment 1, show clear evidence for a dissociation between strength estimates and confidence in those estimates: Ratings of strength and confidence varied in opposite directions when causal power and virtual sample size exerted pressures in opposite directions. For the two extreme conditions across which power increased from .25 to 1.0 as virtual-sample size decreased from 64 to 4, the causal-strength ratings increased significantly (mean difference = 31.9), $F(1, 36) = 26.2$, $MSE = 1{,}473.6$, $p < .001$, $\eta_p^2 = .42$, while confidence ratings decreased (mean difference = 8.8), although they only approached significance, $F(1, 36) = 3.7$, $MSE = 795.0$, $p = .062$, $\eta_p^2 = .09$. Notably, as in Experiment 1, a Measure × Condition analysis of variance revealed that this interaction was highly significant, $F(1, 36) = 20.9$, $MSE = 753.4$, $p < .001$, $\eta_p^2 = .37$. In summary, all results relevant to the conflation hypothesis and to the dissociation between strength and confidence were replicated in Experiment 2.

The close correspondence between the results of Experiments 1 and 2 is striking. The influences of virtual sample size, actual sample size, and causal power are highly reliable for all dependent measures in both experiments. The initial confidence ratings in Experiment 1 correlated with deviations in the causal-strength ratings from causal power predictions in Experiment 2. Moreover, causal strength and sample size influenced strength ratings and confidence ratings in opposite directions in Experiment 2, showing the same pattern of results as in Experiment 1. Thus, it does not appear that the initial confidence question or any of the related procedures compromised the generality of those results that are relevant to the two hypotheses under consideration.

General Discussion

Summary

In Experiment 1, we manipulated actual sample size by varying the number of subjects in hypothetical experiments that had a before-and-after design. We also manipulated the virtual size of each actual sample by varying the probability of the target effect in the before condition. These manipulations led our participants to give systematically different confidence estimates for potential results to be obtained from the hypothetical experiments when they were presented with data from only the before conditions. These initial-confidence ratings, which are unaffected by ΔP because data from the after conditions had not been presented yet, explain the deviations from causal power observed in causal-strength ratings.

Our results also provide clear evidence for a dissociation between estimates of causal strength and confidence in those estimates: (a) A large proportion of participants' strength ratings remained unchanged when virtual sample size changed 16-fold;

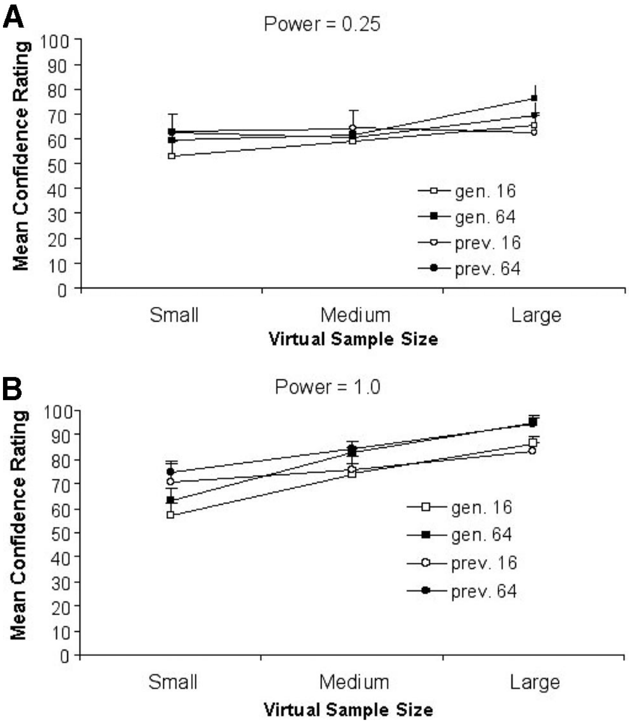

Figure 7. Experiment 2: Mean confidence ratings as a function of virtual sample size with causal direction and actual sample size as parameters, for a power of (A) .25 and (B) 1.0. Error bars represent standard error of the mean. gen. = generative; prev. = preventive.

(b) across conditions at two extremes in which causal power and sample size varied in opposite directions, observed causal-strength and confidence ratings likewise changed in opposite directions; and (c) the strength ratings, but not the confidence ratings, had bimodal distributions, with the primary mode coinciding with causal-power predictions and a secondary mode coinciding with ΔP predictions.

In Experiment 2, we replicated the results of Experiment 1 without the initial confidence question to evaluate the impact of any bias introduced by the use of this novel measure. Notably, every finding in Experiment 1 relevant to our conflation and dissociation hypotheses was replicated in Experiment 2. All conclusions in Experiment 1 were therefore supported by our replication results.

Taken together, these findings clearly refute the weighted-ΔP model (Lober & Shanks, 2000), which can explain neither the influence of actual sample size on strength ratings, the correlation between the initial-confidence ratings and the deviations of participants' strength estimates from causal power predictions, nor the primary mode of the strength ratings coinciding with causal-power predictions. Instead, consistent with the conflation hypothesis, our results suggest that observed causal-strength ratings often deviated from the predictions of causal power because, although subjects estimated causal power, they often conflated their strength estimate with their confidence in the estimate, which clearly varies with virtual sample size.

One important factor to consider in the interpretation of our results is that the trials of a given hypothetical study were summarized and graphically organized according to the presence and absence of the cause and the effect (see Figure 1). This format, in contrast to a sequential and random trial presentation, enables a purer measure of the functioning of the causal learning process itself, relatively free of the distortions to reasoning performance resulting from memory and attentional constraints. In view of the large sample sizes and number of conditions in our experiments, the use of this format is especially important for our goal, that of understanding what the causal-learning process computes.

However, for a different goal, that of describing everyday causal-learning performance, this format may be considered unrealistic. For that context-dependent goal, questions may arise about whether similar results would have been obtained had the trials been presented in a manner more typical of everyday encounters, that is, sequentially and randomly. Previous related research manipulating presentation format suggests that the answer is probably yes. In two of Lober and Shanks's (2000) experiments (Experiments 1 and 3), trials were presented sequentially and randomly, but in two analogous experiments (Experiments 4 and 6), trials were summarized and graphically organized as in our materials. Across presentation formats, when causal power was held constant while ΔP varied, observed strength ratings deviated from causal power in the direction predicted by ΔP, but conversely, when ΔP was held constant while power varied, observed strength ratings deviated from ΔP in the direction predicted by causal power. Thus, the presentation format made no difference to the pattern of results in question.

Causal Support: A Bayesian Decision About the Existence of a Causal Relationship

Now consider the relevance of a recently proposed Bayesian model of causal judgments, the *causal support* model. Tenenbaum and Griffiths (2001) argued that in Buehner and Cheng (1997), participants' causal judgments reflect confidence in a binary decision about the existence of a causal relationship rather than estimates of causal strength. Their Bayesian causal support model (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001) decides whether a set of observations was generated by a causal graphical structure in which a link exists between candidate cause *c* and effect *e* (Graph 1; see Figure 8A) or by a causal structure in which no link exists between *c* and *e* (Graph 0; see Figure 8B).

The causal support model itself might be interpreted as a formalization of the conflation hypothesis (Buehner & Cheng, 1997). However, Griffiths and Tenenbaum (2005, p. 375) explicitly argued against that hypothesis, noting that it implicitly characterizes certainty as being "of secondary importance in evaluating causal relationships" and that "viewing causal induction as a structural inference makes it apparent that neither strength nor reliability should be considered primary: rational causal inferences should combine both factors" (Griffiths & Tenenbaum, 2005, p. 375). In other words, unlike Buehner and Cheng (1997), Tenenbaum and Griffiths do not attribute the conflation of strength and confidence in strength to the ambiguity of the experimenter's question.

A possible interpretation of their approach is that causal strength estimates and confidence in those estimates are indissociable. If they are separable and ambiguity is not an issue, why would

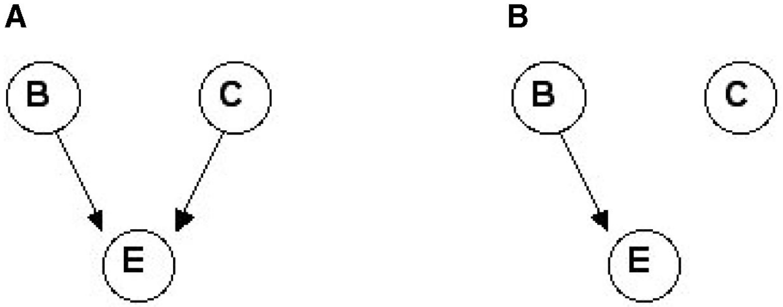

Figure 8. Two candidate graphical structures with a background cause (B), a candidate cause (C), and an effect (E). A: Graph 1, a causal structure in which a link exists between C and E. B: Graph 0, a causal structure in which no link exists between C and E.

participants in Buehner and Cheng's (1997) experiments not have given strength ratings when they were so requested? Our present experiments, using a less ambiguous strength question than that asked by Buehner and Cheng (1997; Lober & Shanks, 2000), provided several lines of converging evidence for participants' ability to dissociate their estimates of strength from confidence in those estimates. (For an analysis of causal-strength ratings in terms of a Bayesian model of strength estimates, as an alternative to one in terms of structure selection, see Lu et al., in press.)

Challenges for the Conflation Hypothesis and Directions for Future Research

Recall that for all three measures in Experiment 1 (initial confidence, strength, and final confidence), the influence of virtual sample size was greater in generative conditions then preventive ones. Moreover, all three measures varied with the ratio of virtual to actual sample size (*R*) in generative but not preventive conditions.[11] These results are not predicted by any model or hypothesis considered in this article and, to our knowledge, they have not been previously demonstrated. As a tentative explanation that may guide future research, we have proposed that a high base rate of *e* signals the presence of alternative generative causes and thus raises doubts about attributing *e* to the candidate cause. In contrast, a low base rate of *e* does not lead to any corresponding doubts due to alternative preventive causes. The absence of *e* is the default state of *e*, which requires no explanation unless there is a change from *e* occurring. Increasing the base rate of *e* therefore reduces confidence for generative causes more than it increases confidence for preventive causes.

Other aspects of our results also appear to challenge the conflation hypothesis. For example, in spite of our efforts to ask a less ambiguous strength question, a large number of participants apparently conflated strength estimates and confidence in those estimates. In contrast, the conflation hypothesis predicts that, given an unambiguous strength question, estimates of strength and confidence in those estimates should be completely dissociable. One possible explanation for this discrepancy is that the strength question posed in Experiments 1 and 2 still retained a fair amount of ambiguity. Specifically, the binary choice format (*absolutely no influence on headache* vs. *produces headache*) might have drawn attention to the degree of confidence in the existence of a causal relationship and might have had residual influences on participants' subsequent numerical estimate.

Participants' interpretation of the strength question may also explain why a small proportion of strength ratings deviated numerically from causal power exactly as predicted by ΔP. Specifically, causal power, when applied to the question of how much difference a cause makes to the probability of an effect in the learning context, would yield ΔP as the answer (Buehner et al., 2003). Of course, these "question interpretation" explanations for why some participants conflated strength and confidence in strength while yet others gave ratings that coincided with ΔP are speculative, and further research is clearly needed to systematically explore the influence of wording on causal judgments.

It is worth noting, however, that both types of deviations from causal-power predictions can be coherently explained under the causal-power approach. If, however, one assumes that all participants compute causal strength as ΔP, one would have no basis for predicting the influence of virtual sample size on the initial confidence ratings; recall that whereas the denominators of the causal power equations define the virtual sample, nothing in the ΔP model restricts causal inference to this subset of the actual sample. Likewise, if one assumes that strength estimates and confidence in those estimates are indissociable, there would be no explanation for why the primary mode of strength ratings coincided with causal power across conditions. Indeed, a major advantage of the causal power approach is that it is compositional; that is, it allows the flexible yet logically consistent formulation of answers to a potentially infinite range of causal questions (e.g., Cheng, 1997, 2000, 2004; Cheng, Novick, Liljeholm, & Ford, 2007; Griffiths & Tenenbaum, 2005; Novick & Cheng, 2004).

[11] Note that our ordinal definition of virtual sample size coincides with *R*. Consequently, the influence of virtual sample size may be due in part to *R*. This possibility, however, does not affect the validity of our discriminating tests of the conflation hypothesis and the weighted-ΔP hypothesis: Regardless of whether variations in the initial confidence ratings are due to virtual sample size or *R,* only the conflation hypothesis predicts that those variations will account for deviations of strength estimates from causal power predictions.

References

Anderson, J. R., & Sheu, C. F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition, 23,* 510–524.

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,* 1119–1140.

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford, England: Clarendon Press.

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

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

Cheng, P. W. (2004). *The deductive nature of causal induction.* Paper presented at the 28th International Congress of Psychology, Beijing, China.

Cheng, P. W., Novick, L. R., Liljeholm, M., & Ford, C. (2007). Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. In M. O'Rourke (Ed.), *Topics in contemporary philosophy: Vol. 4. Explanation and causation* (pp. 1–32). Cambridge, MA: MIT Press.

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

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*: *General and Applied, 79*(1, Whole No. 594)*,* 17.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107,* 195–212.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P.W., & Holyoak, K.J. (in press). Bayesian generic priors for causal learning. *Psychological Review*.

Mish, F. C., et al. (Eds.). (2003). *Merriam-Webster's collegiate dictionary* (11th ed.), Springfield, MA: Merriam-Webster.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455–485.

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.

Shanks, D. R. (2002). Tests of the power PC theory of causal induction with negative contingencies. *Experimental Psychology, 49,* 1–8.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 59–65). Cambridge, MA: MIT Press.

Received April 26, 2006
Revision received June 17, 2008
Accepted August 25, 2008 ■