Supplementary Study

The supplementary study aimed to replicate the central finding from Experiment 2 that participants can use knowledge of the underlying nomological machine to disambiguate confounded regularity information. We addressed two further issues in this experiment. First, in Experiment 2, it was not necessary for participants in the experimental conditions to think through the complete change process taking place within the machine. Rather, it was sufficient to analyze the first part of the device (the chute or the see-saw) in order to get the predictions right, as this part determined the side at which the balloon would appear. In order to demonstrate that participants are in fact capable of using more parts of the device arrangement to make causal predictions, we added in the present experiment a new transfer condition in which the devices in the test phase were different from those observed in the learning phase. The changes targeted later stages of the change process and affected an additional property of the expected outcome (color of the balloon) on the respective side. If participants succeeded in the prediction of both the side and color of the balloon, this would provide evidence that they simulate the complete change process taking place in the device when different start arrangements were realized, rather than simply focusing on the first part. If participants succeeded, this would also provide initial evidence for their ability to transfer knowledge from a learning device to a different transfer device with partly different components.

Second, we wanted to demonstrate that participants readily induce regularities from information about the device arrangement, rather than just making predictions for singular causal events. This would indicate that capacity knowledge supports expectations about law-like dependency relations. To show this, we changed the dependent measure. Instead of gauging predictions for singular unobserved test objects, we asked participants a generic question of how likely it was that different effects would result from throwing different cause objects into the machine, and contrasted these ratings with ratings about how likely the same

effects would be when no objects had been thrown into the machine. The resulting probability increase can be interpreted as reflecting covariation or causal strength intuitions (delta-*P*, see Perales et al., 2017).

Method

Participants.

We recruited 186 participants as in Experiment 2, using the same inclusion criteria and ruling out repeated participation. Twenty-six participants (14%) did not complete the whole survey and were therefore excluded from all analyses. The remaining sample of 160 participants (mean age 33.76, *SD* = 11.19) received £ 0.50 for their participation (corresponding to an estimated hourly wage of £ 6.00, assuming a completion time of five minutes). The sample size was based on the analysis of participants' "specificity scores". The analysis of these specificity scores is the critical test of our hypothesis that people infer and apply capacity knowledge. The relevant conditions for this analysis are shown in
Figure S4 below. Our goal was to have 25 or more participants (after exclusions) in each of these conditions. Our plan was to analyze participants' specificity scores with one-sided one-sample *t*-tests. With 25 participants per condition, the effect size that can be detected with 80% test power is $d = 0.51$. A screenshot of the G*Power sensitivity analyses is provided in the analysis materials that can be downloaded from the repository site.

Design and Procedure.

We implemented a complete factorial 2 (capacity: size vs. weight, between-subjects) × 2 (cause-effect assignment: small & light → left / large & heavy → right vs. small & light → right / large & heavy → left, between-subjects) × 2 (test machine: known vs. novel, between-subjects) × 4 (test object: small & light vs. small & heavy vs. large & light vs. large & heavy, within-subject) mixed design. The basic procedure was the same as in Experiment 2. Participants first learned the effects of two cause objects differing in two confounded features as in Experiment 2. The capacity factor was varied in the learning phase as in the

experimental

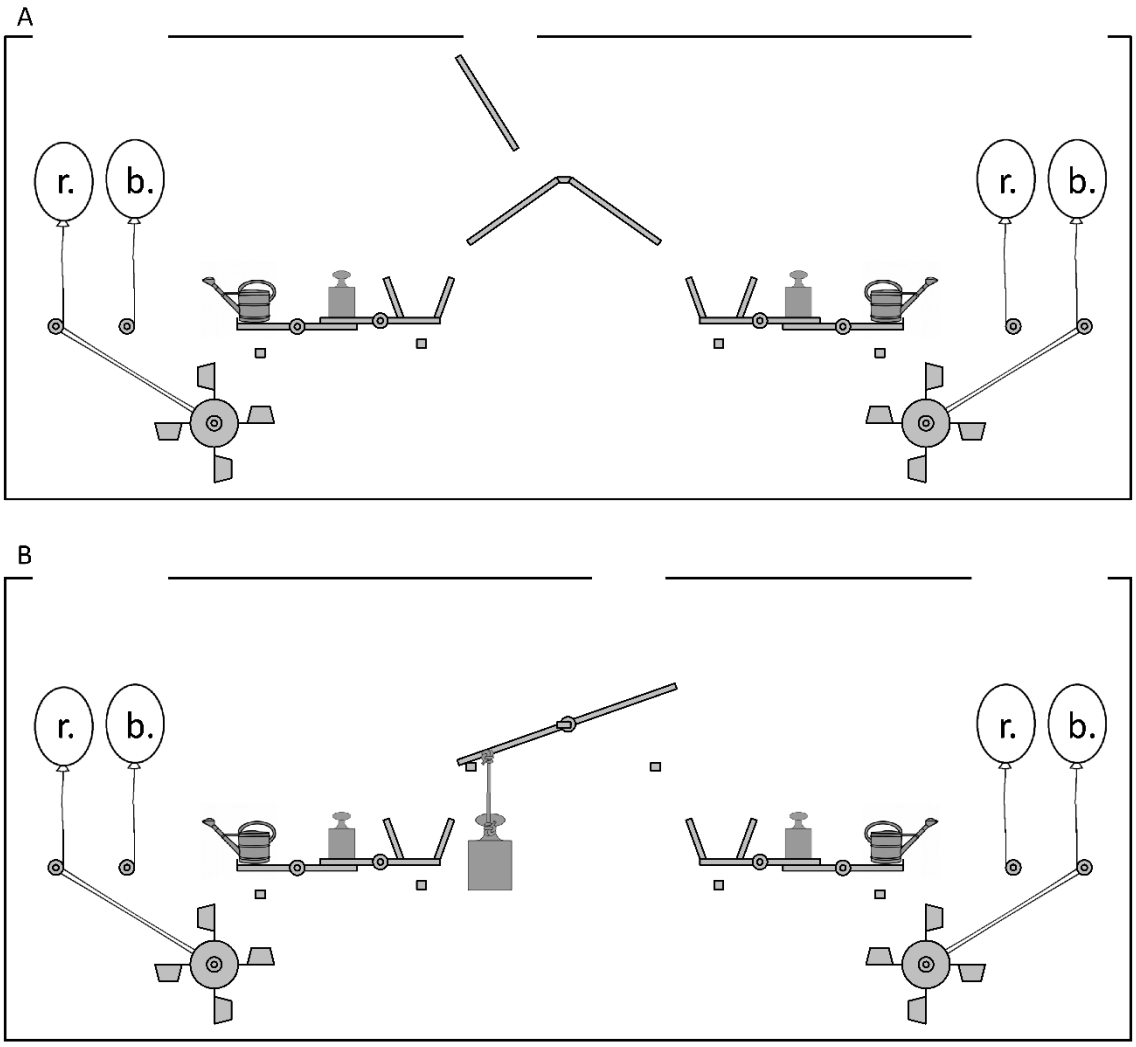

Fig. S1. Construction plans of the *size* condition (A) and the *weight* condition (B), both in the *small & light → left / large & heavy → right* version.

conditions of Experiment 2. (We did not include a control condition without construction plans in this experiment.) The construction plans displayed on the devices in the size vs. weight conditions can be seen in Figure S1. The main difference to the plans in Experiment 2 is that there are two balloons on each side of the machine, one red and one blue, which could potentially rise from the hole above. However, only one of them is attached to the water wheel via a belt, corresponding to the balloon that was observed on the respective side in the learning phase (red on the left side, blue on the right side in all conditions). In both capacity

conditions, we counterbalanced which of the two cause objects led to which of these two effects (factor cause-effect assignment). Construction plans were mirrored accordingly in order to be consistent with the contingency information learned in the animation.

The two *observation* questions after the learning phase included four (rather than two) response alternatives. Participants had to complete the sentences "Throwing the small (3 in) and light (10 lbs..) object [*the large (4 in) and heavy (30 lbs..) object*] into the machine…" with one of the following alternatives: (a) "… caused a red balloon to rise from the left hole", (b) "… caused a blue balloon to rise from the left hole", (c) "… caused a red balloon to rise from the right hole", or (d) "… caused a blue balloon to rise from the right hole", mentioning all four displayed balloons and lowering the probability of chance hits to .25. We excluded participants from the analyses who failed to select the correct answer to any of the two questions.

An additional slide was inserted after the observation questions. Half of the participants were informed that they would "see the same device again" on the next screens and were asked to "tell us what [they] would expect to happen if [they] put different objects into the center hole of the machine". For these participants, the construction plans of the devices displayed in the following test trials were identical to the plan they had seen in the learning phase (condition: test device *known*). This is analogous to Experiment 2, in which participants were also tested on the same device they had previously observed. The other half of the participants (condition: test device *novel*) instead received the following instructions: "On the next screens, you see the same device again, but it has been slightly modified in *two places*. Please have a close look at the new construction plan to find out what the two differences are in comparison to the version you have seen before. Then please tell us what you would expect to happen if you put different objects into the center hole of the new machine" (emphasis in original). In the construction plans presented to these participants in the test phase, the belts on both sides now connected the water wheel with the other balloon compared to the learning

phase (the blue balloon on the left and the red balloon on the right in all conditions; see example in Fig. S2).

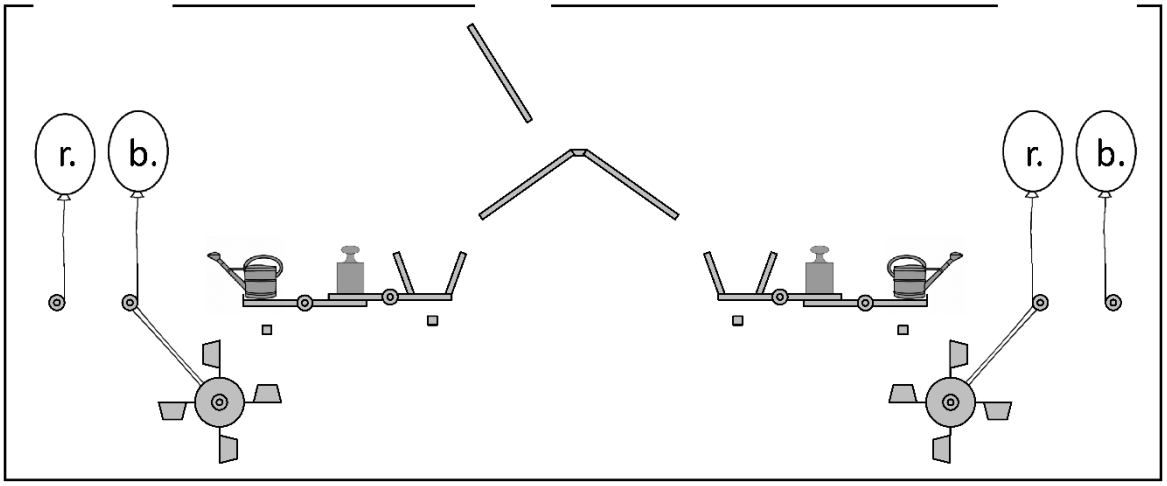

Fig. S2. Example of a modified construction plan shown in the test phase of the condition *capacity: size / test machine: novel*. Water wheels are now attached to the other balloon.

The following test phase also differed in several regards from Experiment 2. There were now five (rather than two) test trials which were administered in random order. In each test trial, four objects with all possible feature combinations were displayed above the machine. On four of the trials, an arrow pointed from one of the four objects (test object: small & light vs. small & heavy vs. large & light vs. large & heavy) into the center hole. Below the illustration, participants read the following question in the version corresponding to the current test object: “Imagine you would throw a small (3 in) [*large (4 in)*] and light (10 lbs.) [*heavy (30 lbs.)*] object into the center hole of the machine. How likely are the following events to happen?" On each trial, this question was followed by the description of four effect events (all combinations of “A red/blue balloon will rise from the left/right hole”), and each description was accompanied by an 11-point scale ranging from 0 (“impossible”) to 100 (“certain”). These four ratings represent the conditional probability estimates for each of the four effects given the presence of the cause object in question on the current trial. On the fifth

trial, a red cross was displayed above the center hole instead of an arrow. Participants were asked to rate the likelihood of the same four events given that nothing was thrown into the center hole of the machine. These four ratings represent the base rates of the four effects in the absence of any cause object. In total, each participant thus gave 20 probability judgments in the test phase. No separate confidence ratings were elicited. Participants were debriefed, thanked, and rewarded as in Experiment 2 after having completed the test phase.

Hypotheses.

The small & light test object and the large & heavy test object were identical to the cause objects observed in the learning phase (henceforth *known objects*), whereas the small & heavy test object and the large & light test object were previously unobserved (henceforth *novel objects*). In essence, the design results in four different inference tasks of increasing complexity. If both the test device and the test objects are known, the task is just to remember which effect had been elicited by the current cause object and to raise the conditional probability of this effect (but not the others) in the presence of this object above its base rate. If the test device is known and the test object is novel, the task additionally requires disambiguating the confounded learning input using the correct capacity cue as implied by the device arrangement. This is analogous to the task in Experiment 2. If the test device is novel and the test objects are known, such disambiguation is not necessary because the effects of these exact objects have already been observed. The side of the correct effect can thus simply be recalled from the learning phase. However, participants in this condition need to realize that the color of the expected effect will change due to the difference in the device arrangement. Correct performance thus requires predicting an as yet unobserved effect on the basis of understanding the underlying nomological machine. Finally, if both the test device and test objects are novel, neither side nor color of the effect event can be directly recalled from the learning phase. Success on this task requires a combination of both previous inference tasks. If participants were able to solve this task, this would constitute evidence that

they do not only consider the first device involved in the change process, but are rather able to simulate the complete change process taking place in the device under different, previously unobserved arrangements. We hypothesized that participants would be able to solve all four tasks, although performance might be somewhat diminished in the more complex tasks that required more transfer.

Data coding and analyses.

As in Experiment 2, we first coded for each participant whether they answered both observation questions correctly. All statistics reported below refer to the subsample that met this requirement; all participants who answered at least one observation question incorrectly were excluded. All analyses were repeated with the whole sample and generally yielded identical conclusions; exceptions are explicitly stated in the text below.

Next, we recoded the four effect events rated on each trial (left/red, left/blue, right/red, right/blue) according to the abstract inference option they represented on each particular trial. Recoding of the side feature looked a bit different for trials with known vs. novel test objects. In cases of known test objects (small & light, large & heavy), we coded for each effect whether it was located on the identical side as the effect that had been elicited by the same cause object in the preceding learning phase ("identical") or on the opposite side ("opposite"). The correct effect is always on the "identical" side. In cases of novel test objects (large & light, small & heavy), we coded for each effect whether it occurred on the same side as the effect that had previously been elicited by a cause object of the same size ("size") vs. weight ("weight"). The correct answer depends on the capacity condition ("size" in the size condition, "weight" in the weight condition). Recoding of the color of the effects was identical for all test objects, as the color depended only on the test device condition. We coded for each effect on each trial whether it had the same color as the effect previously observed on the respective side ("same"), or whether it had a different color ("different"). The

correct answer depends on whether the device is known ("same" is correct) or novel ("different" is correct).

After recoding the 16 cause-effect relationships (each of the four test objects paired with each of the four effects) for each participant, we calculated a covariation estimate (delta-*P*) for each relationship by subtracting each participant's estimate for each effect in the absence of any cause from the same participant's conditional probability estimate of the corresponding effect in the presence of the corresponding cause. These 16 values thus express how strongly the presence of each cause object raises the probability of each of the four effects (see results in Fig. S3).

In a final step, we calculated how strongly each test object raised the probability of the *correct* effect relative to the incorrect effects. To this end, we subtracted the mean delta-*P* estimate for the three incorrect effects from the same participant's delta-*P* estimate for the correct effect on each trial for each participant. The resulting value reflects how specifically participants raised the conditional probability of the correct effect on each trial, compared to the incorrect effects (see results in Fig. S4). These values were subjected to separate one-sample t-tests against 0 for each of the conditions which tests whether participants were able to solve the task in all conditions. Subsequently, we conducted a mixed ANOVA to investigate whether performance was affected by the difficulty of the inference task.

Results

Participants took a median completion time of 348.5 seconds for the task. Thirty-nine out of the 160 participants answered at least one of the observation questions incorrectly and were excluded from the analyses. The delta-*P* estimates calculated from the likelihood ratings by the remaining 121 participants are summarized in Figure S3. A first visual impression confirms that participants in all conditions strongly increased the probability of the correct effect, and increased it much less for the incorrect effects in each condition. For the known objects (averaged across small & light and large & heavy, which yielded highly similar

results) in the known device (Fig. S3a), they predominantly increased the probability of the exact same effect (side and color) that had been elicited by the corresponding cause object in the learning phase regardless of the capacity condition. In the novel devices (Fig. S3c) participants also predominantly raised the probability of an effect that had been elicited by the same object on the identical side, but this time they predominantly raised the probability of the effect with the *different color* (which is correct due to the changed device arrangement in which the belt now connects the water wheel to the other balloon). For the novel objects (averaged across large & light and small & heavy, which were also treated highly similarly) in the known device (Fig. S3b), we replicated the effects from Experiment 2. In the size [*weight*] condition, participants predominantly raised the probability of the effect (side and color) that had previously been brought about by the cause object with the same size [*weight*]. This again shows that participants use capacity knowledge about the device set-up to disambiguate the confounded learning input. Finally, for novel objects in novel devices, participants managed to combine both inference tasks. They again predominantly raised the probability of an effect on the identical side as the effect that had previously been brought about by the cause object with the corresponding size [*weight*], depending on the capacity condition. However, this time, participants predominantly picked the effect with the different color, corresponding to the changed device arrangement. This suggests that participants were able to solve all four tasks. However, it is also apparent that participants distinguished somewhat less clearly between correct and incorrect effects in the more difficult transfer tasks: they tended to raise the probability of the correct effect less strongly and of the incorrect effects more strongly when predictions concerned novel objects and/or novel devices, compared to known objects and known devices (compare Figs. S3b-d with Fig. S3a).

To confirm these impressions statistically, we analyzed the specificity with which participants raised the probability of the correct effect over and above the probability of the incorrect effects. We calculated a “specificity score” for each participant on each trial by

subtracting from the delta-*P* estimate for the correct effect the average delta-*P* estimate for the three incorrect effects. The group means of this specificity indicator are plotted in Figure S4.

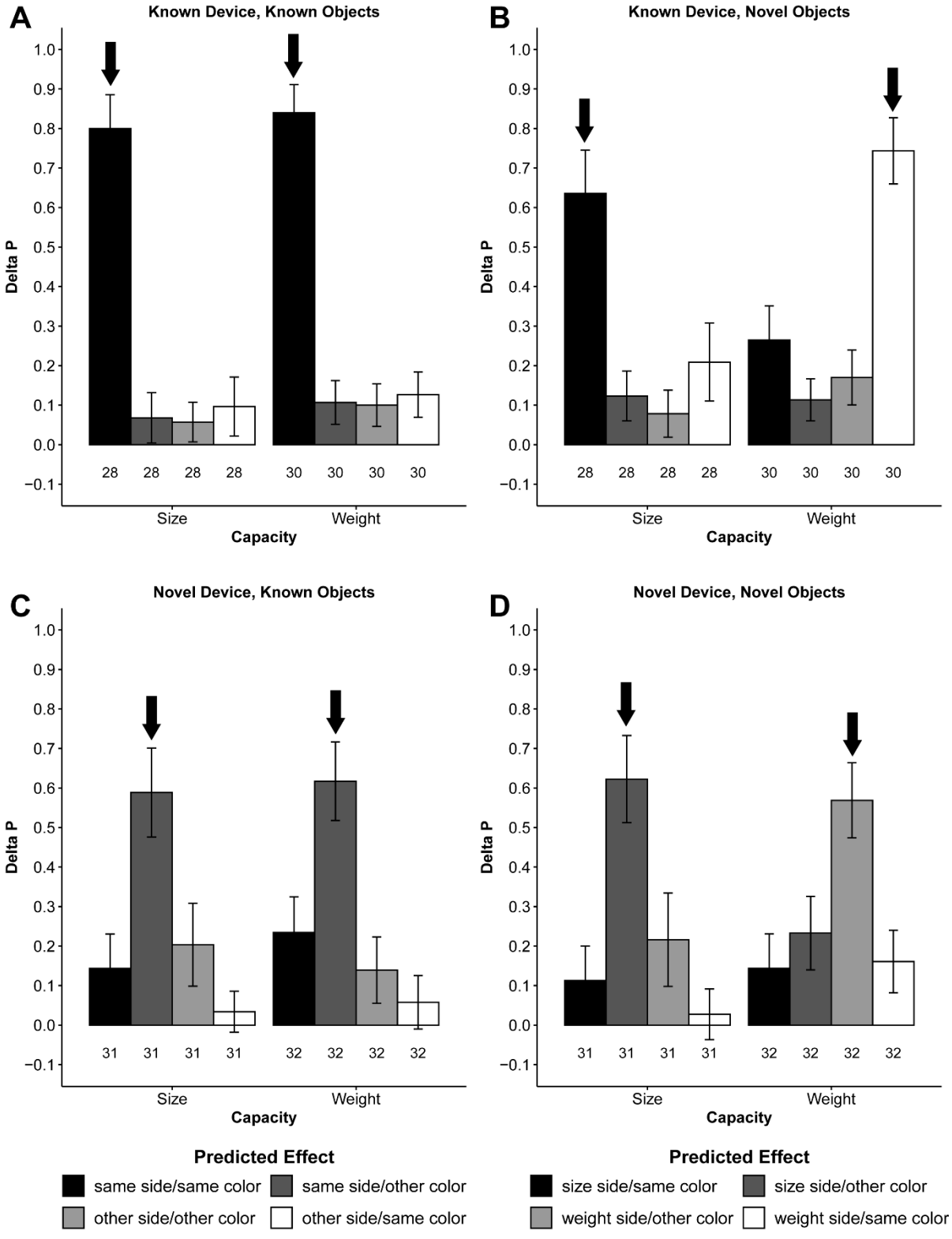

Fig. S3. Delta-*P* estimates from the supplementary study, split up by capacity (size vs. weight, between-subjects), test device (known vs. novel, between-subjects), and type of test object (known vs. novel, within-subject). The bars display condition means of delta-*P* estimates for the four different effects rated on each trial. Error bars represent 95% confidence intervals. The numbers below the bars represent the sample sizes per condition. The correct effect in each condition is indicated by a vertical arrow. The left legend refers to the upper and lower graph on the left side. The right legend refers to the upper and lower graph on the right side.

We assume that if participants had not distinguished the correct from the incorrect effects, they would have changed the correct and incorrect effects from their base rates to the same extent. This would result in a specificity score of 0. We therefore tested the mean specificity score from each condition against 0. All these t-tests were statistically significant, with the largest effect being observed in the known machine / known object / weight condition, $t(29) = 12.29$, $p < .001$, $d = 2.25$, $r = .75$, 95% CI of r [0.52, 0.97], and the smallest effect being found in the novel machine / novel object / weight condition, $t(31) = 4.50$, $p < .001$, $d = 0.80$, $r = .37$, 95% CI of r [0.07, 0.67]. Clearly, participants in all conditions raised the probability of the correct effect more strongly than the probability of the incorrect effects.

Finally, in an exploratory step, we tested whether specificity scores were affected by the inference task. We subjected the specificity scores to a 2 (test machine: known vs. novel, between-subjects) × 2 (test object: known vs. novel, within-subject) × 2 (capacity: size vs. weight) mixed ANOVA. There was a main effect of test machine, $F(1,117) = 5.431$, $p < .05$, $\eta_g^2 = .038$, and a main effect of test object, $F(1,117) = 12.901$, $p < .001$, $\eta_g^2 = .016$. These two main effects were qualified by a significant test device × test object interaction, $F(1,117) = 8.432$, $p < .01$, $\eta_g^2 = .010$, indicating that performance was particularly good for known objects in known test devices (i.e., the pure memory task) and at a relatively constant lower level for the three other trial types. In other words, as soon as at least one type of

inference was required, performance dropped noticeably, regardless of whether it was an inference to a new machine, to a new object, or both. The capacity factor did not affect specificity ratings, nor did it interact with any of the other variables, largest $F(1,117) = 2.301$, $p = .132$ (for the three-way interaction); all other $F(1,117) < 1$. The task difficulty thus seems to be the same for the size-sensitive device (chute) and the weight-sensitive device (see-saw).

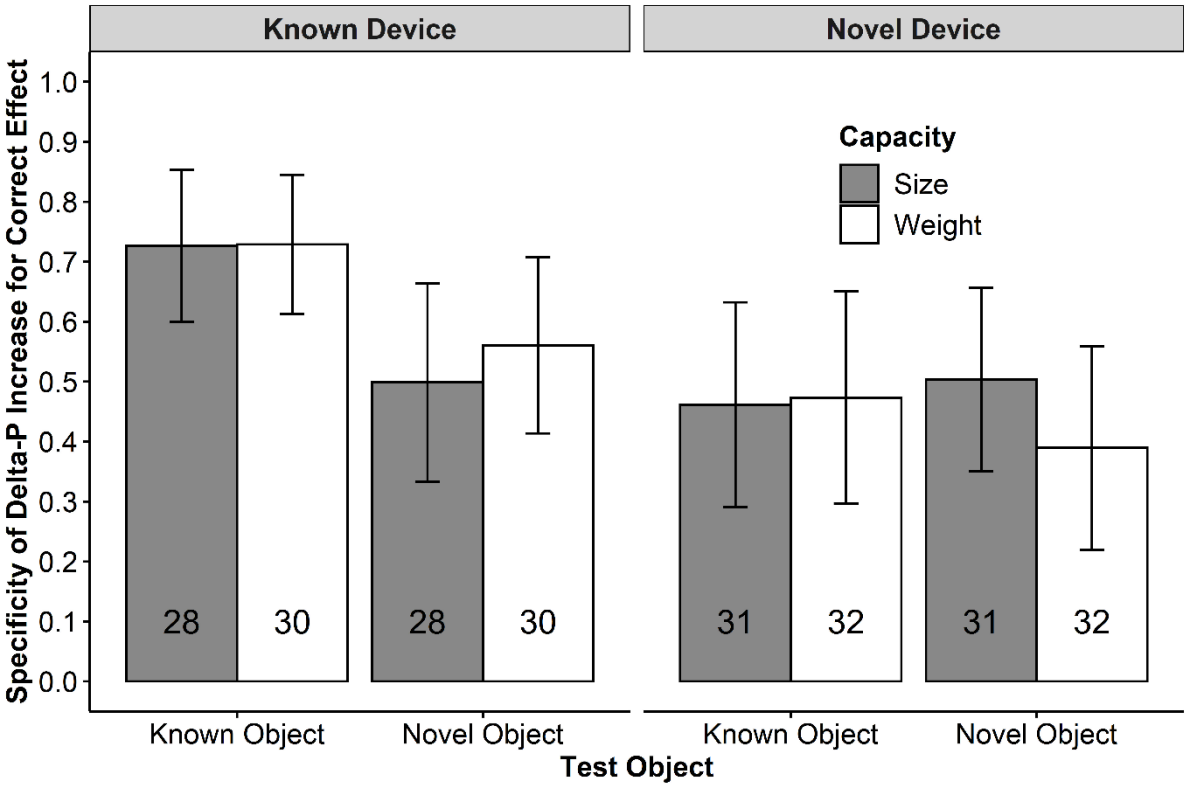

Fig. S4. Specificity of delta-*P* increase for correct effect by capacity, test machine, and test object. Error bars represent 95% confidence intervals. Black numbers in bars represent sample size per condition.

The results of the ANOVA look a bit different if the complete sample is considered that includes the 39 participants who failed in at least one of the observation questions. The specificity ratings are a bit lower overall (ranging from .36 to .62), indicating that these participants introduce more variability to the delta-*P* estimates of the single effects. Specificity is still affected by test object, $F(1,156) = 8.626$, $p < .01$, $\eta_g^2 = .008$, but the effects

of test device and the interaction was non-significant, $F(1,156) = 3.004$, $p = .08$ and $F(1,156) = 2.439$, $p = .12$, respectively. The basic pattern and rank order of the group means still look fairly similar, though, so that we would interpret this difference as mainly due to the unsystematic noise added by inattentive participants. The main conclusion – participants were able to solve all inference tasks but tended to have slightly more difficulties than when simply recalling previously observed cause-effect relationships – can still be drawn even from the complete sample.

Discussion

In the supplementary study, we replicated the finding that participants disambiguate confounded covariation learning input using knowledge of the arrangement of the causal device generating the observed contingencies. They flexibly use capacity knowledge to make predictions for as-yet-unobserved cause objects in new device arrangements. Although performance is slightly reduced when more inferential steps are required in transfer tasks, it is still remarkably high even in the most difficult task. Finally, participants had no difficulty expressing the predictions generated from their capacity intuitions in terms of probabilistic covariation statements.