

# Broken Physics: A Conjunction-Fallacy Effect in Intuitive Physical Reasoning



Ethan Ludwin-Peery<sup>1</sup>, Neil R. Bramley<sup>2</sup>, Ernest Davis<sup>3</sup>,  
and Todd M. Gureckis<sup>1</sup>

<sup>1</sup>Department of Psychology, New York University; <sup>2</sup>Department of Psychology, University of Edinburgh; and <sup>3</sup>Department of Computer Science, New York University

Psychological Science  
1–10

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0956797620957610

www.psychologicalscience.org/PS



## Abstract

One remarkable aspect of human cognition is our ability to reason about physical events. This article provides novel evidence that intuitive physics is subject to a peculiar error, the classic conjunction fallacy, in which people rate the probability of a conjunction of two events as more likely than one constituent (a logical impossibility). Participants viewed videos of physical scenarios and judged the probability that either a single event or a conjunction of two events would occur. In Experiment 1 ( $n = 60$ ), participants consistently rated conjunction events as more likely than single events for the same scenes. Experiment 2 ( $n = 180$ ) extended these results to rule out several alternative explanations. Experiment 3 ( $n = 100$ ) generalized the finding to different scenes. This demonstration of conjunction errors contradicts claims that such errors should not appear in intuitive physics and presents a serious challenge to current theories of mental simulation in physical reasoning.

## Keywords

reasoning, inference, intuitive physics, prediction, open data, open materials, preregistered

Received 8/26/19; Revision accepted 6/22/20

Successful interaction with our environment often requires us to estimate the likelihood of particular physical events. For example, when deciding whether to walk through a construction site, we might gauge the chance of being injured by a falling piece of scaffolding. Accurately assessing the risk requires us to estimate the probability that certain physical events might occur in the future (e.g., a bolt might come loose). Frequently, we might also need to judge the probabilities of the occurrence of conjunctions of several events (e.g., a support tube bends and a bolt comes loose).

There are good reasons to expect that estimates of probabilities in everyday physical situations should be well calibrated and internally consistent. In contemporary simulation theories of physical reasoning (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017), the probability of physical events is argued to be estimated by sampling from noisy simulations consistent with the known state of affairs. A key consequence of this claim is that these estimates should satisfy the constraints of probability theory contingent

on the samples themselves. Finally, a growing body of experiments suggests that humans, even preverbal infants, are sometimes capable of probabilistic reasoning about physical situations (Téglás et al., 2011; Xu & Denison, 2009; Xu & García, 2008).

Further, there are reasons to expect physical reasoning to be different from other forms of reasoning. Every one of our ancestors had to navigate the same physical world, the parameters are exceptionally stable, and correct physical reasoning is particularly valuable in evolutionary terms. For these and other reasons, philosophers and cognitive scientists have argued that intuitive physics will be unlike other forms of commonsense reasoning. Strevens (2013) specifically conjectured, in part based on results with infants, that the well-known flaws that people demonstrate in probabilistic reasoning (cf. Kahneman, Slovic, & Tversky, 1982) appear only in

---

## Corresponding Author:

Ethan Ludwin-Peery, New York University, Department of Psychology  
E-mail: elp327@nyu.edu

dealing with subjective likelihoods (“epistemic probabilities”) and not in dealing with physical probability. In recent work, Firestone and Scholl (2016, 2017) argued that intuitive physics is similar to low-level, automatic perceptual processes. Our aim in this article is to document and report a novel error in reasoning that represents a challenge to this view.

## The Conjunction Fallacy

If a reasoner estimates the probability of physical event A (e.g., a bolt comes loose on some scaffolding) as  $P(A)$  and the probability of physical event B (e.g., a support tube bends) as  $P(B)$ , logically the probability of both events occurring must be equal to or lower than that of either component occurring, that is,  $P(A \wedge B) \leq P(A)$  or  $P(B)$ . However, decades of research have revealed that for many described scenarios, people tend to rate a conjunction as more likely than one or both of its constituents (Tversky & Kahneman, 1982, 1983), a reasoning error known as *the conjunction fallacy*.

The classic article first reporting the conjunction fallacy (Tversky & Kahneman, 1982) included evaluations of a woman named Linda who fit the description of a progressive (e.g., Linda was described as concerned with social justice and in opposition to nuclear weapons). On the basis of this description, participants responded that Linda was more likely to be a feminist than to be a bank teller. Surprisingly, 85% of participants also rated “Linda is a bank teller and is active in the feminist movement” as more likely than “Linda is a bank teller,” a logical impossibility. A common explanation of this error is that the conjunction statement mentions a representative trait or event (being a feminist) that is rated as highly probable on its own, whereas the single trait (being a bank teller) seems less representative of the evoked stereotype.

The conjunction fallacy has been explored in numerous subsequent articles. For example, Tversky and Kahneman (1983) tested several variations on the Linda problem. This included replications with both between- and within-subjects designs and tests on populations with different levels of statistical skill (including undergraduates, medical students, and decision-science PhD students), with all variations confirming the original result. The conjunction fallacy has, of course, received a great deal of theoretical and empirical scrutiny since its introduction (Fiedler, 1988; Gigerenzer, 1991; Hertwig & Gigerenzer, 1999), much of the scrutiny focusing on concerns around pragmatics. However, rigorous empirical work has provided continued support for the finding and its status as a genuine reasoning fallacy (Bonini, Tentori, & Osherson, 2004; Sides, Osherson, Bonini, & Viale, 2002; Tentori & Crupi, 2012).

## Reasoning About Conjunctions of Physical Events

The conjunction fallacy has been examined with a range of different materials, including judgments of the traits of individuals, estimations of the likelihood of natural disasters, predictions about federal legislation, and medical diagnoses (Tversky & Kahneman, 1983). Despite this, to our knowledge it has never been documented in the domain of physical reasoning.

The potential existence of a conjunction fallacy in the domain of physical reasoning poses serious problems for theoretical accounts of physical reasoning based on mental simulation (Battaglia et al., 2013; Ullman et al., 2017). In these models, predictions about the likelihood of various events depend on examining one or more outcomes of a mental simulation that maintains an approximate isomorphism to the physical dynamics of the actual world, similar to the way video games approximate real physical dynamics. Such theories suggest that to estimate whether a tower of blocks will fall over, reasoners form an approximate mental representation of the configuration of each block in the tower, run forward a number of mental simulations each from a slightly different starting point (owing to sources of perceptual uncertainty about the precise configuration of the starting state of the simulation), and make final judgments by aggregating over the results of these simulations.

Probabilities estimated using this type of Monte Carlo simulation necessarily conform to the axioms of probability theory. For example, the frequency with which A and B both occur in different randomly initialized simulations must be less than or equal to the frequency of either event occurring alone across those simulations. There is no way for A and B to occur without A occurring as well, meaning judgments made using relative counts across a sample of simulations to estimate probabilities will always avoid the conjunction fallacy. This consistency with the laws of probability is a key virtue of the simulation approach, enabling sophisticated forms of inductive inference (Ullman et al., 2017). Thus, irrespective of accuracy, these theories predict that there should be no systematic violation of the axioms of probability in subjective judgments about common physical scenarios.

Often, it is assumed that the classical theory of probability is the correct method of representation, but there are other theories of probability. A body of recent work has suggested that cognitive models of judgments may be better fitted by the more general quantum-probability framework (Pothos & Busemeyer, 2013; Pothos, Busemeyer, Shiffrin, & Yearsley, 2017). These data may be relevant to this debate as well.

## Experiment 1

To evaluate the possibility of a conjunction fallacy in the domain of physical prediction, we employed a within-subjects design in which each participant viewed a number of clips showing simple physical scenes in a 2-D world. Participants viewed the first few seconds of each scene and rated the probability of a future event occurring if the scene were to continue (e.g., “What is the probability the ball will fall in the hole?”). Rating the probability of specific future events is a common task that has been used in many recent studies on intuitive physical reasoning (Battaglia et al., 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Hamrick, Smith, Griffiths, & Vul, 2015). Each critical scene appeared twice, but participants were not informed of this fact. For these eight critical scenes, participants rated a conjunction event  $P(A \wedge B)$  on one appearance and one constituent event  $P(A)$  on the other. If participants rate the conjunction probability as more likely than the constituent probability, this is a form of the conjunction fallacy.<sup>1</sup>

## Method

**Participants.** We recruited 90 participants (28 female; age:  $M = 33.6$  years,  $SD = 9.8$ ) on Amazon’s Mechanical Turk. Of these participants, 74 were able to answer basic comprehension questions about the task, given three attempts. Sixty-two participants were eligible for our analysis on the basis of the exclusion criteria (outlined below). We analyzed only the first 60 participants (18 female; age:  $M = 34.2$  years,  $SD = 9.7$ ), as stated in our preregistration (<https://osf.io/gvknw>). Pilot testing of the materials suggested that the effect was robust and could be reliably detected even with small sample sizes; on the basis of the pilot data, we chose this sample size to provide high power ( $> .95$ ) to detect the effect.

**Materials and procedure.** After accepting the survey and consenting to participate, participants read a detailed description of the task. This included several example videos of the physics engine we used<sup>2</sup> and example clips such as those that appeared in the main body of the survey. These examples included many forms of interobject interactions, including collisions, and participants were allowed to watch these videos as many times as they wanted. Participants were informed of the nature of the clips, and we explained how we wanted them to report their estimates of likelihood. Participants then answered seven simple comprehension questions about the task and were given three attempts to do so. If they were able to answer these questions correctly, they moved on to the rest of the experiment.

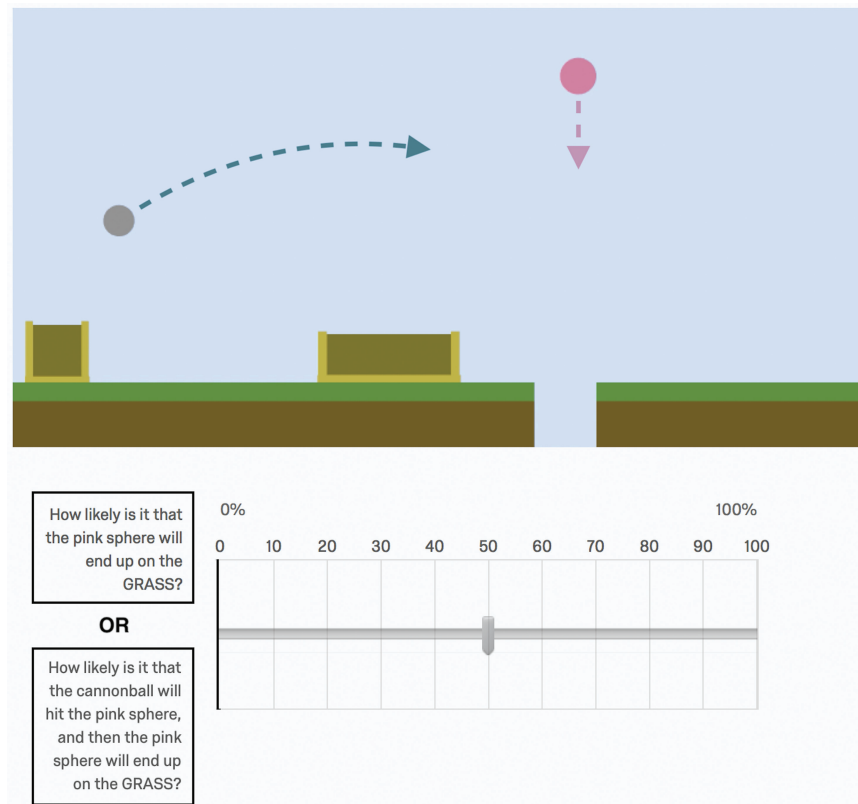
In the main body of the experiment, participants saw several scenes in which a pink “sphere” dropped toward a hole in a grassy field, and a gray “cannonball” traveled across the scene in such a way that it could potentially collide with the pink sphere (Fig. 1). One object was called a cannonball and the other a sphere so that participants would be less likely to confuse them given the written description. Pilot testing indicated that calling both objects by the same name caused confusion.

There were several minor differences among the scenes as well, including the exact speed and position of the objects, the size of the hole, and the presence or absence of one or more boxes on the grass. Each video stopped after approximately 700 ms, well before the cannonball could possibly intersect with the pink sphere’s path, leaving ambiguity about the outcome of the scene. For each scene, participants were asked to estimate the likelihood of a particular outcome and express that estimate as a percentage chance.

Eight of the scenes were critical, the answers to which provided our primary measure. Each critical scene appeared twice, but participants were not informed of this fact. Half of the scenes that appeared twice were mirrored horizontally in their second appearance. For each scene that appeared twice, in one appearance participants were asked the question, “How likely is it that the pink sphere will end up on the GRASS?”<sup>3</sup> and in the other, “How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?” Scenes did not repeat until after several filler scenes were presented and completed. All video materials are available on OSF (<https://osf.io/jsqpd>).

For the filler scenes, participants were asked questions unrelated to the outcomes of interest used in the critical trials, such as the likelihood that the cannonball might end up in one of the boxes or in the hole. Pilot data indicated that separation by a few filler scenes was sufficient to prevent participants from explicitly recognizing the repetition.

Following the completion of the main body of the experiment, participants were asked to describe how they answered the questions in the main task using the following prompt: “Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below.” We also asked several open-ended questions intended to determine whether or not participants had noticed that some of the scenes appeared twice with different questions. Finally, participants answered several demographic questions, gave free-response feedback, and were debriefed. (All data and materials are available on OSF at <https://osf.io/jsqpd>.)



**Fig. 1.** An annotated example of a scene and its associated questions shown to participants in Experiment 1. Dotted arrows indicate approximate motion over the approximately 700-ms-long movie clip. The gray circle was described as a “cannonball” and the pink circle as a “sphere.” Each scene appeared twice, once with the constituent question, “How likely is it that the pink sphere will end up on the GRASS?” and once with the conjunction question, “How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?”

## Results

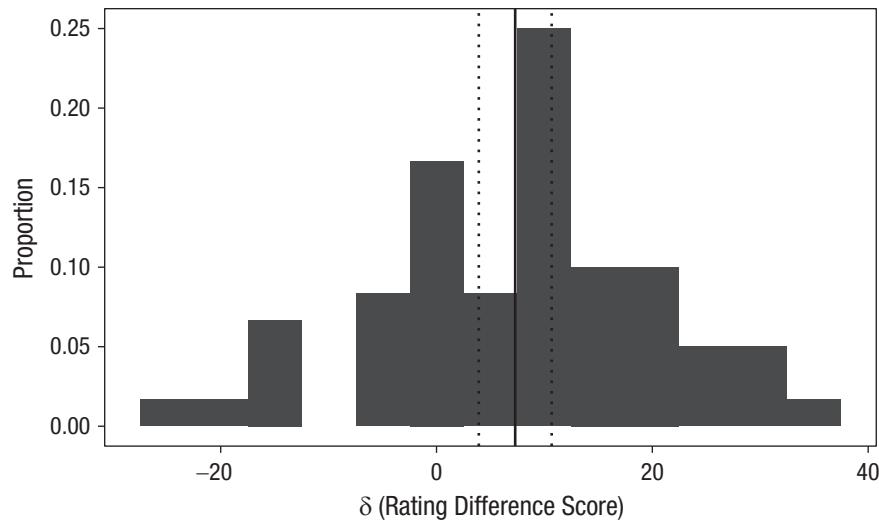
The experiment included two questions with trivially obvious outcomes (e.g., the cannonball had already missed the sphere and could not possibly collide with it). When answering these questions, 12 participants (16%) reported that the near-certain outcome was less than 90% likely or that the near-impossible outcome was more than 10% likely; these participants were not included in our analysis, as defined in our preregistration.

**Primary analyses.** For each of the eight critical problems, participants rated the probability of the conjunction and sole statement. The mean judgment of the conjunction across all problems was 44.69% ( $SD = 25.88$ ), and the mean judgment of the sole event across all problems was 37.40% ( $SD = 26.58$ ).

The difference between these two ratings formed our primary data. We averaged the rating difference scores ( $\delta = \text{conjunction rating} - \text{sole rating}$ ) for each participant

for each of the eight problems (Fig. 2). Positive values of  $\delta$  indicate that participants rated conjunctions as more likely than their constituent sole events, which is a form of the conjunction fallacy. Zero or negative scores are not fallacious. The average  $\delta$  value was 7.29% ( $SD = 13.07$ ,  $SE = 1.69$ ), which was reliably greater than zero, according to both a two-tailed one-sample  $t$  test,  $t(59) = 4.32$ ,  $p < .001$ , and one-sample Bayesian estimation (BEST; Kruschke, 2013), 95% credible interval (CrI) = [4.06, 10.79]. The effect size ( $d$ ) was 0.56, 95% confidence interval (CI) = [0.28, 0.83]. Therefore, participants appear to have erroneously rated conjunctions as more likely than their constituent sole events.

As shown in Figure 2, when we averaged all critical trials together, 72% of participants rated the conjunctions as more likely than their constituents on average. In addition, 62% of participants rated the conjunction as more likely than the constituent on more than half of the critical pairs. If participants were respecting the laws of probability in their estimations, they would generally



**Fig. 2.** Distribution of the average of the eight difference scores for each participant (Experiment 1). The solid vertical line indicates the mean, and the dotted vertical lines indicate the 95% confidence interval.

rate the conjunction as less or equally likely. We would certainly not expect to see a consistent reversal.

**Secondary analyses.** In the postexperiment questionnaire, none of the participants reported noticing that some of the videos appeared twice. In a follow-up question revealing that some of the videos appeared twice, only seven participants claimed to notice. The questionnaire also asked participants to estimate how many videos were repeated. Although the true number of repeats was eight, only three participants guessed close to this number. The majority of participants who guessed said that only a small number (two or three) were repeated, although many participants declined to guess at all.

At the end of the experiment, we asked participants, “Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below.” Three coders who were not involved in the design or running of the experiment or the collection of data were asked to code the free responses into four categories. The ratings had a Cronbach’s alpha of .75, indicating acceptable agreement (Kline, 2013). A one-way, between-subjects analysis of variance (ANOVA) found a significant effect,  $F(3, 56) = 3.900, p = .013, \eta^2 = .17$ , and a Bayesian test produced a Bayes factor (BF) of 4:1 in favor of a difference by reported approach. However, because this result failed to replicate in the higher powered Experiment 2, we suggest that it was a false positive and do not draw any conclusions from it.

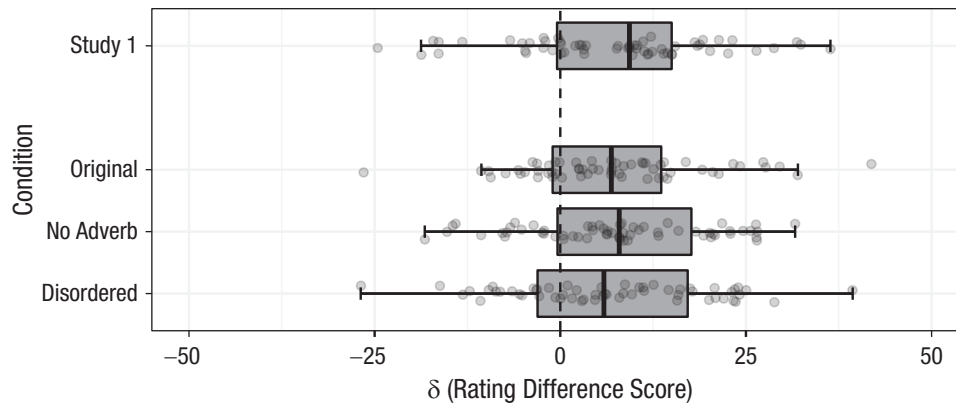
## Experiment 2

Experiment 1 provided evidence that people commit the conjunction fallacy when reasoning about simple

physical scenarios. However, there is reason to be cautious in interpretation. Early criticisms of conjunction-fallacy results centered on the idea that participants might be interpreting the task in line with conversational norms (Gigerenzer, 1991, 1996). People expect statements in conversation to be informative, truthful, relevant, and clear (Grice, 1991), but it can be argued that conjunction questions sometimes violate these expectations, leading participants to answer a slightly different question from the one intended. Various empirical attempts have been made to demonstrate that apparent conjunction-fallacy errors are simply the result of participants reasonably misinterpreting the materials (Dulany & Hilton, 1991; Fiedler, 1988; Hertwig & Gigerenzer, 1999; Mellers, Hertwig, & Kahneman, 2001; Morier & Borgida, 1984), although even in the strictest tasks, the errors persist (Sides et al., 2002; Tentori & Crupi, 2012).

Two alternative interpretations seem particularly of concern in Experiment 1. Mellers et al. (2001) pointed out that, for example, “We invited friends and colleagues to the party” implies a union of friends and colleagues, rather than an intersection. If participants are reading the question quickly, they could potentially interpret  $P(A \wedge B)$  as something like  $P(A \vee B)$ , that is, “How likely is it that the cannonball will hit the pink sphere *or* the pink sphere will end up on the grass *or* both events will occur?” Similarly, participants might interpret the conjunction as the conditional (“*If* the cannonball hits the pink sphere, *then* how likely is it that the pink sphere will end up on the GRASS?”). The conditional being larger than one constituent is not a logical impossibility and certainly not a form of the conjunction fallacy.





**Fig. 3.** Distribution of the average of the eight difference scores for the participants in each group (Experiment 2). Data from Experiment 1 are included for comparison. Circles indicate individual participants. The vertical line in each box is the median, and the ends of the boxes correspond to the first and third quartiles. The whiskers extend to the farthest point that is less than 1.5 times the interquartile range from the box ends.

To account for these classes of alternative interpretations, we ran a replication of Experiment 1 with additional conditions, allowing us to evaluate whether alternative phrasings of the conjunction question would lead to the same conjunction errors observed in Experiment 1.

## Method

**Participants.** We recruited 269 participants (98 female; age:  $M = 35.3$  years,  $SD = 10.2$ ) on Amazon’s Mechanical Turk. Exclusion criteria were the same as in Experiment 1. We analyzed only the first 180 participants (60 female; age:  $M = 36.2$  years,  $SD = 10.2$ ), as stated in a new preregistration (<https://osf.io/ga98v>). This sample size was chosen so that each of the three conditions would have the same sample size as Experiment 1—60 participants after exclusions.

**Materials and procedure.** Experiment 2 was identical to Experiment 1, save for the phrasing used in the appearances of the conjunction question. We tested three different phrasings of the conjunction question in three between-subjects conditions. The first was the original-phrasing condition, in which participants saw the same phrasing that appeared in Experiment 1, namely, “How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?”

We compared this with two alternative phrasings. The first was based on the original phrasing but omitted the connecting adverb *then*. As a result, we call this the no-adverb condition. The comma was similarly omitted. The conjunction question in this condition was, “How likely is it that the cannonball will hit the pink sphere and the pink sphere will end up on the GRASS?” The final condition was phrased to highlight the purely conjunctive nature of the question being asked. No causal language was used, and the components were

presented in the reverse of what one would expect to be the natural order of events, so we call this the disordered condition. The conjunction question in this condition was, “How likely is it that both will happen: The pink sphere will end up on the GRASS and the cannonball will hit the pink sphere?”

Each participant was randomly assigned to one of the three conditions. Whenever they saw a conjunction question, they were given the appropriate phrasing for that condition. (All data and materials are available on OSF at <https://osf.io/jsqpd>.)

## Results

The experiment included the same two trivially obvious questions as in Experiment 1. When answering these questions, 36 participants reported that the near-certain outcome was less than 90% likely or that the near-impossible outcome was more than 10% likely; these participants were not included in the analysis.

**Primary analyses.** The mean judgment of the conjunction across all problems was 46.53% ( $SD = 28.20$ ), and the mean judgment of the sole event across all problems was 39.35% ( $SD = 28.78$ ).

As in Experiment 1, we averaged the rating difference scores ( $\delta$ ) for each participant for each of the eight problems (Fig. 3). Across all three conditions, the average rating difference score was 7.18% ( $SD = 12.34$ ,  $SE = 0.92$ ), which was reliably greater than zero, according to a two-tailed one-sample  $t$  test,  $t(179) = 7.81$ ,  $p < .001$ , and a one-sample BEST, 95% CrI = [5.30, 9.00]. The effect size ( $d$ ) was 0.58, 95% CI = [0.42, 0.74]. Again, participants systematically rated conjunctions as more likely than a constituent sole event.

The effect of condition on rating difference scores was not significant,  $F(2, 177) = 0.068$ ,  $p = .93$ ,  $\eta^2 = .001$ ,

and Bayesian analysis found strong evidence of no difference (BF favoring the alternative over the null hypothesis [BF<sub>10</sub>] = 0.067). We dummy-coded condition in a linear regression and in a Bayesian linear regression, the original condition being used as the reference level, to compare the new phrasings with the original phrasing. Neither slope was significantly different from zero, all  $ps > .80$ , all Bayesian 95% CrIs for the slopes including zero, indicating no differences between each alternative condition and the original condition.

None of the 95% CIs on the slopes indicated differences that would reduce any condition to zero. Further one-sample  $t$  tests showed consistent differences from zero,  $p < .001$  for all three conditions, and all Bayesian 95% CrIs not including zero.

**Secondary analyses.** As before, three new coders coded free-response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of .78, indicating acceptable agreement (Kline, 2013). A chi-square test found no evidence of a relationship between condition and approach used,  $\chi^2(6, N = 180) = 9.56, p = .14$ , and a Bayesian test of association produced a BF of 5:1 against a relationship between condition and approach, suggesting that the different phrasings did not influence choice of the approach used to solve the problems.

To determine whether there was an overall impact of approach, we conducted a one-way between-subjects ANOVA to compare the effect of reported approach on overall ratings on the critical items. There was no significant effect,  $F(3, 176) = 0.582, p = .63, \eta^2 = .01$ , and a Bayesian test produced a BF of 13:1 against a difference by reported approach.

### Experiment 3

Experiment 2 established that the physical conjunction-fallacy effect is highly consistent across alternative phrasings of the critical question, suggesting that the physical scenario itself, rather than the pragmatics of the question, produces this judgment pattern. However, Experiments 1 and 2 investigated only one specific instance of the physical conjunction fallacy (two balls colliding in midair). One concern is that some unexplored idiosyncrasy of our original design was responsible for this effect. By exploring whether we would find this error in a range of more and less similar scenes, we rounded out the evidence for a physical conjunction fallacy. To that end, in Experiment 3, we designed and tested a range of different physical scenarios that had varying similarity to those in Experiments 1 and 2.

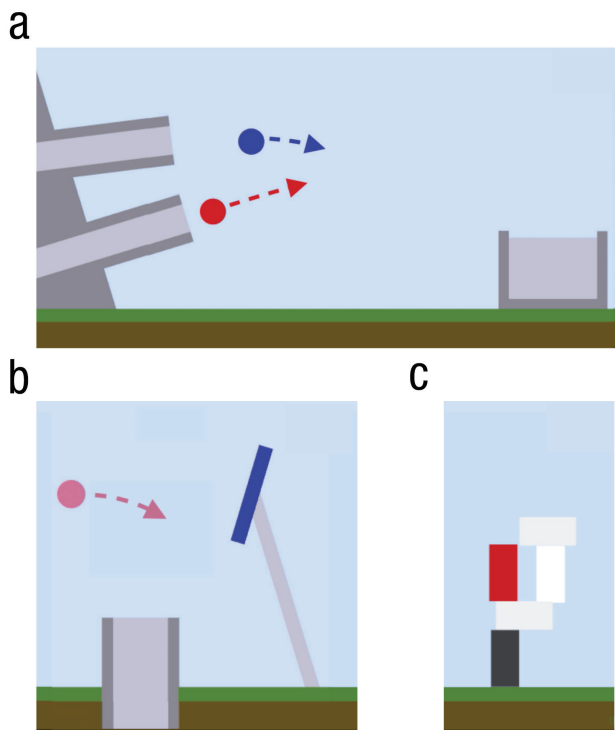
### Method

**Participants.** We recruited 198 participants (70 female; age:  $M = 36.89$  years,  $SD = 10.66$ ) on Amazon's Mechanical Turk. Exclusion criteria were the same as in Experiment 1. We analyzed only the first 100 participants (36 female; age:  $M = 38.34$  years,  $SD = 10.17$ ), as stated in a new preregistration (<https://osf.io/pyb8q>), because this new experiment was added at the request of reviewers.

**Materials and procedure.** We pilot-tested a few different types of physical scenes before settling on three designs to develop further. These three new types of scenes were developed to investigate how widely the physical conjunction fallacy appears. As a result, the first new type of scene is somewhat similar to the scene in our original design, the second departs in certain critical ways, and the third is intentionally different along several axes. All materials are available on OSF (<https://osf.io/jsqpd>).

The first scene, "tubes," is moderately similar in design to the original scenes (Fig. 4a). Like the scenes used in the previous experiments, tubes scenes involved two balls, both moving. As in our original scenes, one of the key questions is about whether or not the two balls will collide, and the other question concerns where one of the balls will end up. This type of scene always involved a red ball and a blue ball, each exiting a tube, moving in the same direction, and flying toward a bucket, which the balls might land in but might equally overshoot. The sole question was always, "How likely is it that the BLUE ball will end up in the BUCKET?" and the conjunction question was always, "How likely is it that the RED ball and the BLUE ball will collide, and the BLUE ball will end up in the BUCKET?"

The second scene, "basket tube," represents more of a departure from our original design (Fig. 4b). Although these scenes also included a possible collision event, there was only ever one moving object, a single pink ball. This limited the amount of information that participants needed to keep track of to make an informed decision. The object in motion and its trajectory would always be the focus of attention, so different ways of asking the question should not bring new sets of objects into scrutiny. Each scene showed the pink ball flying in an arc above a field with a gray tube sticking up out of it. Each scene paused at a point where the ball clearly would not fall into the tube if continuing on its parabolic trajectory. However, there was always a blue "backstop" ahead of the ball, which the ball could possibly hit. If the ball were to hit this backstop, it was possible that it might bounce off the backstop and into the tube. The sole question was always, "How likely is



**Fig. 4.** An annotated example of the three new scene types used in Experiment 3, approximately as they appeared to participants. Dotted arrows indicate approximate directions of motion, if any. All scenes were, in actuality, the same dimensions and are truncated here to conserve space. An example of a tubes scene (a) is shown at the point where the video clip stopped. An example of a basket-tube scene (b) is shown at the point where the video clip stopped. An example of a still-weight-tower scene is shown in (c). In this type of scene, the tower was presented as a still image rather than as a short video clip.

it that the ball will end up in the TUBE?” and the conjunction question was always, “How likely is it that the ball will hit the BLUE backstop and the ball will end up in the TUBE?”

The final type of new scenes, “still weight tower,” represents the greatest departure from our original design and therefore shows the greatest generalization of the effect (Fig. 4c). These scenes were still images, rather than short videos; they involved no objects with initial motion; they involved the complex interaction of more than two objects, rather than simple collisions; and although in all other designs, both components of the conjunction were events, in this case one of the critical questions was about a parameter (the weight) of a particular object in the scene. Each scene portrayed a small standing tower of five or six blocks. Some blocks were pale, some were dark, and one block was always bright red. To orient participants to the scene, we told them, with every such image, “The DARK blocks are HEAVY, the PALE blocks are LIGHT, and the RED blocks might be either HEAVY or LIGHT.” As part of the design, these red blocks of ambiguous weight

were always placed at a location in the tower where the question of their weight might contribute seriously to the tower’s overall stability. The exact sole and conjunction questions differed somewhat between the different towers, but the sole question always concerned whether or not the tower would stay standing (e.g., “How likely is it that the tower will STAY STANDING?”), and the conjunction question always added a question about the weight of the red block (e.g., “How likely is it that the RED block is HEAVY and the tower will STAY STANDING?”).

There were four scenes of each type, for a total of 12 new scenes. Each of these scenes appeared twice, once with the conjunction question and once with the sole question. Scenes were intermixed with a small number of filler questions to help prevent recognition of previous questions. Materials were presented as in the previous experiments. In the main body of the experiment, all participants saw both questions for all of the scenes and estimated the likelihood of the stated outcome, as in previous experiments.

Finally, participants answered several demographic questions, gave free-response feedback, and were debriefed. (All data and materials are available on OSF at <https://osf.io/jsqpd>.)

## Results

Experiment 3 included the same two trivially obvious questions as in Experiments 1 and 2. When answering these questions, 10 participants reported that the near-certain outcome was less than 90% likely or that the near-impossible outcome was more than 10% likely; these participants were not included in the analysis.

As in Experiments 1 and 2, we averaged the rating difference scores ( $\delta$ ) for each participant for each of the four problems in each of the three new problem types. As a result, we ended up with a  $\delta$  for each of the new scene types. Positive values of  $\delta$  indicate that participants rated conjunctions as more likely than their constituent sole events, which is a form of the conjunction fallacy.

**Tubes.** The mean judgment of the conjunction across all four scenes of this type was 43.15% ( $SD = 19.04$ ), and the mean judgment of the sole event across all four scenes of this type was 33.07% ( $SD = 19.00$ ). The average  $\delta$  value for the tubes problems was 10.08% ( $SD = 17.22$ ,  $SE = 1.72$ ), which was reliably greater than zero, according to both a two-tailed one-sample  $t$  test,  $t(99) = 5.85$ ,  $p < .001$ , and a one-sample BEST, 95% CrI = [6.40, 13.00]. The effect size ( $d$ ) was 0.59, 95% CI = [0.37, 0.80].

**Basket tube.** The mean judgment of the conjunction across all four scenes of this type was 34.73% ( $SD = 16.10$ ),



and the mean judgment of the sole event across all four scenes of this type was 31.03% ( $SD = 13.80$ ). The average  $\delta$  value for the basket-tube problems was 3.70% ( $SD = 10.93$ ,  $SE = 1.09$ ), which was reliably greater than zero, according to both a two-tailed one-sample  $t$  test,  $t(99) = 3.39$ ,  $p = .001$ , and a one-sample BEST, 95% CrI = [1.60, 5.90]. The effect size ( $d$ ) was 0.34, 95% CI = [0.14, 0.54].

**Still weight tower.** The mean judgment of the conjunction across all four scenes of this type was 44.77% ( $SD = 13.34$ ), and the mean judgment of the sole event across all four scenes of this type was 40.00% ( $SD = 14.85$ ). This is strong evidence that participants were not misinterpreting the conjunction as the conditional. Conditional on the red block's weight being known, the overall likelihood would be much greater than 50%, which we did not observe here. The average  $\delta$  value for the still-weight-tower problems was 4.77% ( $SD = 14.38$ ,  $SE = 1.44$ ), which was reliably greater than zero, according to both a two-tailed one-sample  $t$  test,  $t(99) = 3.31$ ,  $p = .001$ , and a one-sample BEST, 95% CrI = [1.80, 7.40]. The effect size ( $d$ ) was 0.33, 95% CI = [0.13, 0.53].

**Summary.** Overall, it appears that there was a reliable tendency to make conjunction-fallacy errors in all three of the new scene types.

## General Discussion

This article reports three experiments showing that people rate conjunctive events as more likely than their constituents across a set of physical-reasoning problems. Experiment 1 demonstrated the effect, Experiment 2 showed that the effect was robust to a range of alternative phrasings, and Experiment 3 expanded the finding to a wider range of physical scenes.

Although the conjunction fallacy is well established, the detection of a similar effect in physical reasoning is unexpected for several reasons. Many arguments have been made that intuitive physical reasoning is distinct from other types of cognitive activities and will be immune to these types of errors. In addition, common explanations evoked for the conjunction fallacy seem hard to apply in this case. For example, the concept of "representativeness" seems less relevant to the physical domain because there is not such a salient category or schema to activate.

One possibility is that the conjunction fallacy is a general phenomenon that occurs across many domains because it is a fundamental error in our judgment capacities. However, in follow-up work, we have found that although magnitudes of conjunction-fallacy errors are often correlated with one another (e.g., the size of an individual's error on a "Linda" problem correlates with conjunction-fallacy errors on reasoning about

dice), physics conjunction errors do not correlate with these other problems, suggesting a distinct and novel mechanism (Ludwin-Peery, 2020).

We argue that these results are additionally intriguing because they are unexpected given recent accounts of probabilistic mental simulation. However, such theories might be modified in light of these results. For example, rather than aggregating across multiple simulation runs to make a probabilistic inference, people might use some type of biased aggregation scheme that results in judgment errors (Zhu, Sanborn, & Chater, 2020). It remains to be seen whether this biased aggregation approach can provide a simultaneous account of all the other documented phenomena in intuitive physical reasoning.

Importantly, it has frequently been acknowledged in past work that there might be cases in which heuristics were employed instead of mental simulation (Battaglia et al., 2013). If simulation is abandoned for certain problems, it suggests a control problem for the brain to determine when to adopt a simulation and when to use a heuristic. The present experiments provide an important waypoint about when simulation might be abandoned that may help inform such theories.

## Transparency

*Action Editor:* Timothy J. Pleskac

*Editor:* D. Stephen Lindsay

*Author Contributions*

E. Davis conceived the study. All the authors designed and planned the experiments. E. Ludwin-Peery created the materials and conducted the experiments. E. Ludwin-Peery analyzed the data with feedback from T. M. Gureckis. E. Ludwin-Peery wrote the manuscript in consultation with N. R. Bramley, E. Davis, and T. M. Gureckis. All the authors approved the final manuscript for submission.

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.


*Open Practices*

Data and materials for all three experiments have been made publicly available via OSF and can be accessed at <https://osf.io/jsqpd>. The experiments were preregistered on OSF—Experiment 1: <https://osf.io/gvknw>, Experiment 2: <https://osf.io/ga98v>, Experiment 3: <https://osf.io/pyb8q>. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



## ORCID iDs

Ethan Ludwin-Peery  <https://orcid.org/0000-0001-9505-1983>

Neil R. Bramley  <https://orcid.org/0000-0002-4141-8476>

## Acknowledgments

The authors thank Ellie Robbins, Michael Lepori, Xuechen Sheryl Zhang, and Adi Kwiatek for help with this research and Gregory L. Murphy and Gary F. Marcus for helpful discussion.

## Notes

1. A limited account of the results of Experiment 1 was previously published as a conference paper (Ludwin-Peery, Bramley, Davis, & Gureckis, 2019).
2. We used the PhysX physics engine through the Unity interface (<https://unity3d.com>).
3. Possible resting-state locations were presented in all caps for clarity.

## References

- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences, USA, 110*, 18327–18332.
- Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language, 19*, 199–210.
- Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition, 9*, 85–110.
- Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*, 123–129.
- Firestone, C., & Scholl, B. (2016). Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision, 16*(12), Article 689. doi:10.1167/16.12.689
- Firestone, C., & Scholl, B. (2017). Seeing physics in the blink of an eye. *Journal of Vision, 17*(10), Article 203. doi:10.1167/17.10.203
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases.” *European Review of Social Psychology, 2*, 83–115.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592–596.
- Grice, H. P. (1991). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition, 157*, 61–76.
- Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In D. C. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 866–871). Austin, TX: Cognitive Science Society.
- Hertwig, R., & Gigerenzer, G. (1999). The “conjunction fallacy” revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*, 275–305.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kline, P. (2013). *Handbook of psychological testing*: Abingdon, England: Routledge.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*, 573–603. doi:10.1037/a0029146
- Ludwin-Peery, E. (2020). Limited domain structure for conjunction errors. In S. Denison., M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 2710–2716). Montreal, Quebec, Canada: Cognitive Science Society.
- Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2019). Limits on the use of simulation in physical reasoning. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 707–713). Montreal, Quebec, Canada: Cognitive Science Society.
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*, 269–275.
- Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin, 10*, 243–252.
- Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences, 36*, 255–274.
- Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General, 146*, 968–987.
- Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition, 30*, 191–198.
- Strevens, M. (2013). *Tychomancy*. Cambridge, MA: Harvard University Press.
- Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*, 1054–1059.
- Tentori, K., & Crupi, V. (2012). On the conjunction fallacy and the meaning of *and*, yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition, 122*, 123–134.
- Tversky, A., & Kahneman, D. (1982). Judgments of *and* by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge, England: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*, 649–665.
- Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97–104.
- Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, USA, 105*, 5012–5015.
- Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review, 127*, 719–748. doi:10.1037/rev0000190