AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 9, 2636–2650
https://doi.org/10.1037/xge0001399

Do Striking Biases in Mass Inference Reflect a Flawed Mental Model of Physics?

Alex Mitko and Jason Fischer
Department of Psychological and Brain Sciences, Johns Hopkins University

To engage with the physical world, we rely on our intuitive sense of how objects will behave when we act on them or they interact with each other. Objects' latent properties such as mass and hardness determine how their physical interactions will unfold, and people have a keen ability to infer these latent properties by observing physical events. For example, we can precisely discriminate the relative masses of two objects when we see them collide. However, such inferences are sometimes subject to marked biases. When inferring mass from an observed collision, people consistently overestimate the mass of an incoming object that strikes a stationary one. Why? A number of plausible accounts have been put forward, variously arguing that the bias arises from rule-based reasoning, oversimplified stimuli, or noisy perceptual estimates of the scene dynamics. The implications of these views stand in stark contrast to one another: systematic biases may reveal a fundamental deficiency in the mental model of physical behavior, or they may be an expected consequence of reasoning over imperfect information. Here, we investigated all three accounts within a unified paradigm, presenting videos of real-world bowling ball collisions. We found that using richly detailed stimuli did not eliminate biases in mass inference. However, individual differences in the biases were task-specific and well-explained by noisy perceptual estimates rather than oversimplified physical inference mechanisms. Our findings collectively point toward an intuitive physics system that implements Newtonian principles but is subject to the quality of the information it operates on.

Public Significance Statement

People are able to estimate how much an object weights by watching how it behaves—for example, a heavy object will launch a lighter one when the two collide. Although people sometimes make errors when judging objects' weights, our study shows that these errors are not due to an inadequacy in the "physics engine in the mind." Rather, the errors reflect reasonable best guesses when the available information is limited.

Keywords: mass inference, intuitive physics, physical properties, individual differences, heuristics

Imagine your next trip to the grocery store. Perhaps on your way through the packaged foods, you pick up a bag of rice noodles and a jar of garlic ginger sauce. Without giving it much thought, you gently pick up the noodles—they are light and fragile—but use a firm grip on the heavy sauce jar so it will not slip out of your hand. In the produce section, you take a watermelon from the bin and decide against placing it on the bottom rack of the cart where it could easily roll off. As you pack your shopping bags, you pack the milk carton separately from the tomatoes so its sharp corners will not cause any damage, and you opt to carry the tray of deviled eggs by itself, knowing that it will spill if tilted too much. Although you probably were not pondering any formal physics equations while shopping, your implicit understanding of how objects will behave when they interact—termed *intuitive physics*—guided your decisions and actions on the trip (Fischer, 2020; Fischer & Mahon, 2021; Kubricht et al., 2017). In the scenario here, the items' physical properties such as mass, hardness, and smoothness were key to how you interacted with them, and might have even helped you find them more quickly to begin with (Guo et al., 2020).

A critical component of intuitive physics is learning and representing the latent physical properties of the objects around us. For new objects that we encounter, we discover their physical properties in a variety of ways. Perhaps most obviously, we learn about objects by interacting with them—for example, lifting, squeezing, and poking to sample their properties (Lederman & Klatzky, 1987). But our everyday environments contain many more objects than we can personally interact with, and we also learn about latent physical properties

This article was published Online First May 4, 2023.

Alex Mitko https://orcid.org/0000-0002-6437-0006

All of the data presented in this article can be found in raw, deidentified form at https://osf.io/wd6rj. The stimuli for all of the tasks are available at https://www.dynamicperceptionlab.com. A subset of the findings in this article was previously presented at the Vision Sciences Society annual meeting (held virtually) in May 2021 and described in several academic talks.

This work was funded by internal Johns Hopkins University funds.

Alex Mitko served as lead for data curation and investigation. Jason Fischer served as lead for supervision. Alex Mitko and Jason Fischer equally contributed to conceptualization, methodology, software, visualization, writing–original draft, writing–review and editing, and formal analysis.

Correspondence concerning this article should be addressed to Alex Mitko, Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, United States. Email: amitko1@jhu.edu

through observation. For example, people can infer the weights of unfamiliar objects by watching others interact with them (Runeson & Frykholm, 1981; Zhang et al., 2019), or by seeing two or more objects interact with each other. In fact, this latter case formed the basis for some of the early research on intuitive physics, which asked observers to judge the relative masses of two objects that they saw collide and bounce off of one another (Todd & Warren, 1982). The change in the objects' velocities reveals their relative masses in a straightforward way from the standpoint of classical mechanics, but observers' judgments often show surprising patterns. For example, in the case where an incoming object strikes a stationary one, people consistently overestimate the mass of the incoming object relative to the stationary one (Figure 1a; Gilden & Proffitt, 1989; Runeson et al., 2000; Sanborn, 2014; Todd & Warren, 1982). A long line of work has since sought to understand why this bias and other similarly puzzling ones arise in mass judgments, and what these biases mean about the architecture of the intuitive physics system in the mind (Flynn, 1994; Gilden & Proffitt, 1989, 1994; J. B. Hamrick et al., 2016; Runeson et al., 2000; Runeson & Vedeler, 1993; Sanborn, 2014; Sanborn et al., 2013; Todd & Warren, 1982).

Do systematic errors in mass inference point toward a flawed or oversimplified mental model of physical behavior? An early account put forward by Gilden and Proffitt (1989) proposed that people use rule-based heuristics to make their judgments rather than applying Newtonian principles to model the behavior of a scene (Figure 1b). For the mass inference task described above, Gilden and Proffitt proposed that people rely on two rules to make their decisions: (a) the object with a slower postcollision speed is the heavier of the two and (b) if the incoming object ricochets backward after it strikes the stationary object, then the stationary object is the heavier one. The rules agree with each other in most cases, and in fact, the second rule would always lead to the correct response. However, when the two objects are very similar in mass, the presence or absence of a ricochet can be hard to discern because the incoming object is nearly motionless after the collision. In these cases of fine mass discriminations, observers must rely primarily on the first rule, but these are also exactly the cases in which the first rule often fails. Even when the incoming object is lighter, its speed will be slower after a collision with an object close in weight, leading to the conclusion that it is heavier. Together, these rules could produce a pattern of performance like that in Figure 1a.

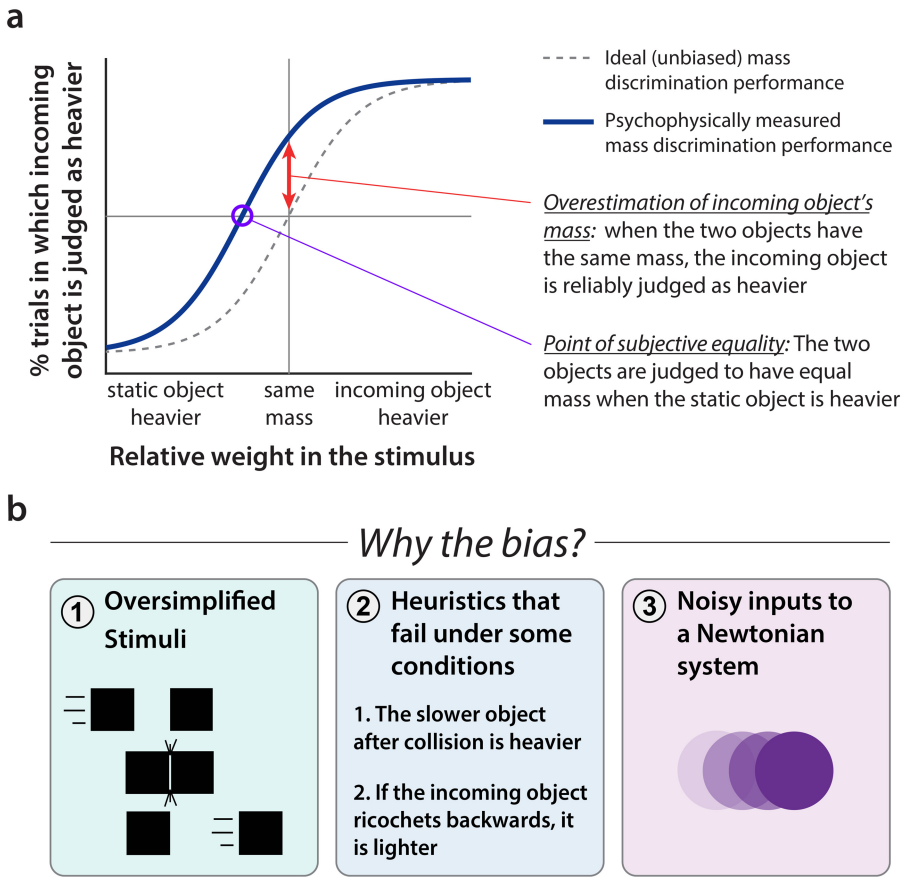
On its face, a heuristic account of the sort put forward by Gilden and Proffitt would seem to provide a parsimonious explanation for an otherwise puzzling pattern of errors in people's physical judgments. Other heuristics might underlie the common misconceptions that heavy objects fall faster than light ones (Vicovaro et al., 2019) and that an object dropped by a moving person will fall straight down (McCloskey et al., 1983). A heuristic approach to physical reasoning could provide a suitable parsing of the physical structure and behavior of the environment most of the time while bypassing the need for a more general mental representation of Newtonian laws. On the other hand, the simplicity with which individual heuristics can be expressed belies the substantial challenge of handling the immense variety of physical scenarios that we encounter in daily life with a piecemeal collection of rules. Properties like friction, elasticity, and three-dimensional geometry that are key determinants of physical behaviors would need to be captured (either implicitly or explicitly) in the rules if they are not represented as variables to which Newtonian principles are applied. Any adequate library of heuristics would be expansive, yet still fall short in scenarios like the mass judgments described above. By contrast, a compact set of Newtonian principles would provide a general way to accurately capture the behavior of the world in any scenario we might encounter. If we could apply some general laws of physics to our observations of the world, they would provide a far more robust way of predicting its behavior to guide our decisions and actions.

This reasoning has led to a recent revival of the notion that people have an implicit understanding of Newtonian laws and apply them to predict how physical dynamics will unfold (Battaglia et al., 2013; Ullman et al., 2017). Accompanying work has provided an alternative account of the biases and apparent misconceptions that are sometimes observed in physical judgments: the accuracy of our physical judgments is constrained by the fidelity with which we can perceive and represent the relevant physical variables for the scenario (K. A. Smith & Vul, 2013). When operating on noisy or incomplete information, even a perfect mental model of physics would yield imperfect predictions. One example of work supporting this notion comes from Flynn (1994), who posited that the constraint in the mass inference task was in the stimulus itself—the lack of real-world cues that are necessary for accurate physical inference. Though many studies have replicated the bias in mass judgments described above, the effect was absent when Flynn (1994) used videos of real collisions as stimuli. Unlike previous studies that used computer-generated movies of shapes colliding, Flynn filmed videos of controlled bowling ball collisions. Participants no longer displayed a bias toward overestimating the weight of the incoming ball, suggesting that the bias was not the result of errors in physical inference per se, but rather the result of oversimplified stimuli. In other work, the use of realistic or interactive stimuli has revealed proficiency in a number of intuitive physics tasks where simpler stimuli had yielded systematic errors (Kaiser et al., 1986; K. A. Smith & Vul, 2013). It remains an open question what exactly is conveyed by realistic stimuli that are absent from more simplified ones, but it appears that real-world behaviors sometimes run counter to the errors predicted by the heuristics model.

More recently, an alternate account was put forward by Sanborn et al. (2013) and Sanborn (2014). They showed that biased judgments can arise from a Newtonian model of physical behavior when the perceptual estimates fed to the model are noisy. Their model, which implemented correct Newtonian physics in a Bayesian framework, revealed a bias in mass inference closely matching that of human judgments, and arising from priors on the velocities of objects. Further, when they occluded the velocity cues that the heuristic account relies on, human and model judgments still successfully discriminated the objects' weights and displayed the same bias that had been previously reported.

In both Flynn's and Sanborn's accounts, there is no need to invoke deficiencies in physical inference mechanisms to account for subjects' biased reports. Rather, these accounts acknowledge that physical judgments are susceptible to imperfect information, and errors manifest in a way that appears puzzling on its surface. Still, we are left with the challenge of reconciling the collective findings of this prior work. Do the rich visual details in videos of real-world collisions reduce perceptual uncertainty to the point that mass inference biases are eliminated as well? Might performance when judging richer stimuli be more in line with the predictions of a heuristic model? A challenge in synthesizing the findings of the prior work outlined here and arriving at a unified account of the findings

Figure 1
Bias in Mass Inference From Observed Collisions



Note. (a) Past research has identified a systematic bias in how people perceive the relative weights of objects after seeing them collide. When a moving object strikes a stationary one, observers reliably overestimate the mass of the incoming object relative to the initially stationary one. The bias can be observed as a shift in the psychometric function relating the relative weights of the two objects in the stimulus to the likelihood that observers will rate the incoming object as heavier: when the two objects are equal in weight, the incoming object is more often judged as heavier; in order for the two objects to generally be perceived as having equal weight, the stationary object must be heavier. (b) A number of potential explanations for the bias have been put forward. The *oversimplified stimuli* account proposes that the bias is a result of presenting diagrammatic animated displays, and is attenuated or absent from judgments in real-world scenarios (Flynn, 1994). *Heuristic* accounts propose that the bias from the application of efficient rules to understand collision dynamics–heuristics support fast and accurate inferences in most scenarios but fail under certain conditions (Gilden & Proffitt, 1989). The *Noisy Newtonian* account proposes that people reason about collision dynamics using an accurate mental model of physical behavior. Noisy perceptual estimates of critical scene contents such as objects' velocities can lead to systematic biases in physical inferences (Sanborn et al., 2013). See the online article for the color version of this figure.

comes from the variety of tasks, stimuli, and analyses that the studies have employed. In the present work, we aimed to disentangle the competing accounts by testing them under common conditions. We introduce a new set of stimuli—video recordings of real-world collisions with a fine-grained manipulation of the objects' relative masses—and present data from a large sample of online participants. This approach provided a sensitive way to measure precision and bias in mass inference both at the group level and within individual participants. Our findings show that providing rich visual cues does not eliminate mass inference biases. In fact, the vast majority of participants showed biases in the direction predicted by prior work, but there were highly reliable individual differences in bias as well. Our follow-up experiments showed that mass inference bias was task-specific and not linked to intuitive physics performance more generally, and that the bias could be well-explained by noisy velocity estimates even though more velocity cues were available than in typical simplified displays. Our findings point toward a robust Newtonian model of physics in the mind, with errors on the mass inference task arising from noisy inputs to the model rather than limitations in the model itself.

Results

Experiment 1

We first tested whether observers displayed the previously reported bias in mass inference (Gilden & Proffitt, 1989; Runeson et al., 2000; Sanborn, 2014; Todd & Warren, 1982) when viewing videos of real-world collisions (Figure 2a). We presented the movies to 85 online participants and tested for a systematic overestimation of the incoming ball's mass. Each participant viewed trials containing all possible pairs of bowling ball weights (balls weighed from 6 to 16 pounds in increments of 2 pounds) and, after seeing the balls collide, indicated which of the two they thought was heavier. We also presented participants with silhouette versions of the videos, created by annotating the videos on a frame-by-frame basis and re-rendering them in silhouette form (see "Material and Method"; Figure 2b). These silhouette videos contained the same translational motion profiles as the full videos but lacked other visual cues such as surface texture and shading. The blocked design alternated between the full videos of the bowling balls and silhouette versions, and performance on the full videos and the silhouette videos was analyzed separately.

Figure 2c shows performance for the two stimulus sets, plotted as the proportion of the time that participants reported the incoming ball as heavier as a function of the relative weights of the incoming and static balls. Logistic curves fit to the group data revealed a significant shift in the point of subjective equality for the full videos (PSE = −0.11; 95% CI = [−0.143, −0.08]; $p < .0001$), indicating that the two balls were perceived as equal in weight when the static ball was actually heavier, indicating an overestimation of the incoming ball's weight. A bias in the same direction was present for the silhouette stimuli (PSE = −0.18; 95% CI = [−0.22, −0.15]; $p < .0001$), and a permutation test revealed that the bias was significantly greater for the silhouette videos ($z = 2.67$, $p < .01$).

These results establish two key findings regarding the source of the bias in participants' mass inferences. First, the significant difference in the measured bias between the silhouette and full video conditions shows that simplifications in the stimulus *do* contribute to the magnitude of the bias, as suggested by the work of Flynn (Flynn, 1994). Second, the fact that a significant bias remained even when participants viewed videos of real-world collisions shows that the use of simplified stimuli is not the sole reason that the bias emerges—even with rich visual details available, participants still reliably overestimate the mass of the incoming ball. Taken together, these findings raise the possibility that the bias has multiple sources that can contribute independently to the overall bias measured within a given paradigm. For example, when presented with a simplified display, people might reason about the properties of the diagram itself rather than the hypothetical real-world scenario it refers to (Schwartz, 1995), using a qualitatively different approach to predict the behavior of diagrams versus real-world objects. In our present data, that might mean that the larger bias measured with the silhouette videos reflects a contribution of diagram-based reasoning that is absent from participants' judgments for the full videos. However, it is also possible that the bias we measured in the two conditions has the same underlying cause, which is exacerbated by the loss of some visual details in the silhouette videos. Perceptual uncertainty would be a prime candidate—if noisy velocity estimates give rise to the bias as proposed by Sanborn et al. (2013)and Sanborn (2014), the larger bias for the silhouette videos could reflect greater uncertainty in the absence of cues from the bowling balls' surface textures and other scene details. We next investigated these two possibilities by characterizing individual differences in performance on the tasks and assessing the degree to which the measured biases are related across tasks.

To characterize individual differences in performance, we fit separate psychometric curves to each participant's data (Figure 3a). The family of curve fits is shifted leftward from the origin (as expected from the group data), and participants vary in the magnitude of their biases. To assess whether these individual differences in bias were reliable, we conducted a split-half analysis, separately estimating the bias in independent halves of each subject's data (see "Material and Method"). We did the same for participants' sensitivity (d-prime) in their weight discrimination, computing d-prime for each independent half of the data. Figure 3b and 3c show the reliability of individual differences in bias and d-prime, respectively. The split-half correlation was strong and significant in both cases ($z = 0.81$, $r = .67$, $p < .001$ for the bias; $z = 1.33$, $r = .87$, $p < .001$ for d-prime).

These reliable individual differences provided an opportunity to test whether participants' biases were correlated across the two independent tasks, which would suggest a common source. We found that individual differences in both bias and d-prime were strongly and significantly correlated across tasks ($z = 0.82$, $r = .68$, $p < .001$ for the bias; $z = 1.42$, $r = .89$, $p < .001$ for d-prime; Figure 3d), which favors the second possibility outlined above—that the bias arises for the same reason in the two tasks and is exacerbated by the removal of real-world visual cues.

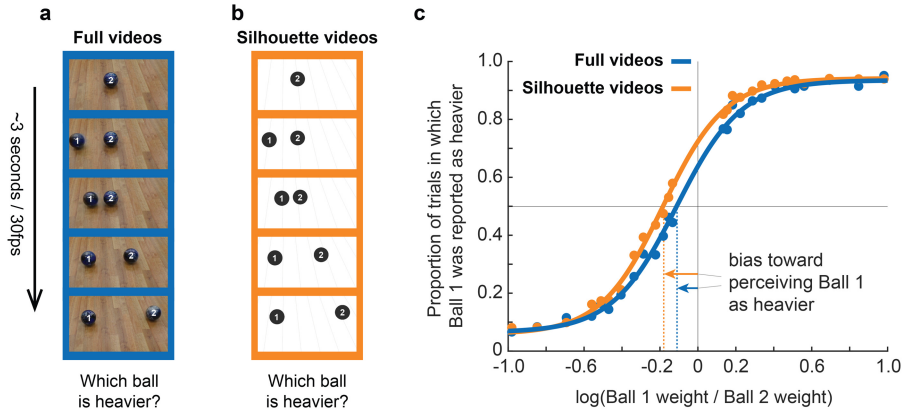
Experiment 2

In addition to establishing that a robust bias in mass inference persists even when observers make judgments about real-world collisions, Experiment 1 also uncovered reliable individual differences in the degree to which participants express this bias. There were reliable individual differences in mass discrimination (d-prime) as well, and together these individual differences provide a powerful tool for examining whether the mass inference bias arises from a fundamental computation within the intuitive physics system. To the extent that other intuitive physics tasks can also capture reliable variation across individuals, a comparison of individual differences across tasks can establish whether an individual's degree of bias in mass inference judgments predicts their performance when reasoning about other physical scenarios. We would expect to observe such a relationship if the mass inference bias arises from a mental computation that is a fundamental component of the intuitive physics system, recruited to reason about other physical events beyond the collision scenarios tested in Experiment 1. To give a hypothetical example of one such possible computation, suppose that the estimation of mass is conflated with momentum in the mental representations of moving objects such that faster-moving objects are treated as though they have larger masses. This would account for the mass inference biases found in Experiment 1 in a straightforward way, and it would also lead to errors in predicting postcollision dynamics in other scenarios where moving objects collide. Individual differences in the degree of misestimation would manifest across many tasks beyond just the bowling ball collisions studied in Experiment 1.

In Experiment 2, we tested the relationship between mass inference bias and performance on other intuitive physics tasks using

Figure 2

Experiment 1: A Persistent Bias in Mass Inference Even When Viewing Real-World Events



Note. (a) Full videos of bowling ball collisions. Each movie clip depicted a rolling ball (Ball 1) entering the frame and striking a stationary one (Ball 2; labels are depicted here for illustration but were not displayed during the experiment). Videos showed the time period from just before Ball 1 entered the frame to the time at which one ball exited the frame after the collision; approximately 3 s. (b) Silhouette videos of bowling ball collisions. Each full video was converted to a silhouette version via frame-by-frame labeling of the balls' positions. Silhouette videos depicted the bowling balls as black circles, and simple perspective cues were added by tracing lines on the ground surface. In all other respects, the silhouette videos were identical to the full videos. (c) Psychometric curve fits to the group data from the full videos (blue) and the silhouettes (orange). Both conditions exhibited a significant bias in the inferred masses of the bowling balls, evidenced by a shift in the point of subjective equality. The bias was significantly stronger for the silhouette videos as compared with the full videos, but a marked bias was still present for the full videos despite their inclusion of rich visual detail. See the online article for the color version of this figure.

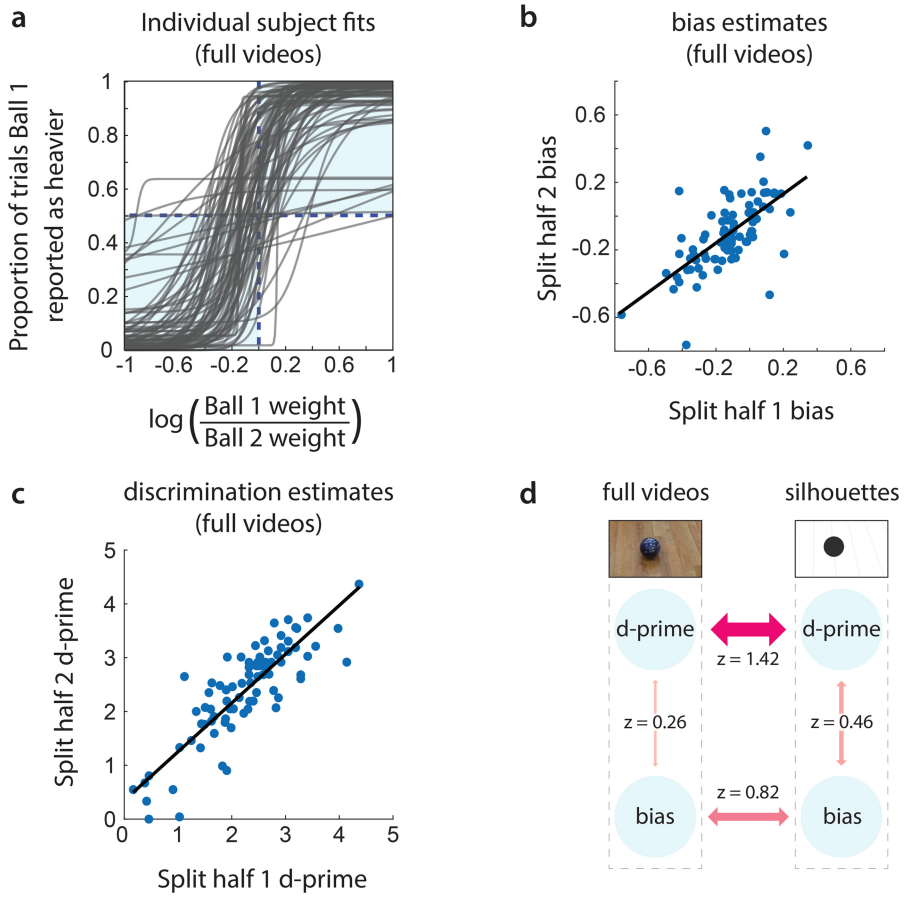
an individual differences approach. Along with completing the Bowling Balls task with the full videos from Experiment 1, participants completed two additional intuitive physics tasks (Figure 4): the toppling towers task (Figure 4a), which we have previously shown to yield reliable individual differences in performance (Mitko & Fischer, 2020) and the bouncing discs task (Figure 4c), based on one used by Fischer et al. (2016). Participants also completed the mental rotations test (Peters et al., 1995; Shepard & Metzler, 1971; Vandenberg & Kuse, 1978), which was included to control for other factors that could drive individual differences in the aforementioned tasks such as spatial ability or level of motivation. In the toppling towers task, participants judged how unstable block towers would fall and predicted where the majority of the blocks would ultimately come to rest (gray or white side of the platform). In the bouncing discs task, participants saw two discs moving within an arena, and were instructed to track the predicted trajectory of one of the discs for a period of 2 s when it was rendered invisible. When the disc reappeared, participants indicated whether it appeared in the correct location and with the correct velocity, as though it had continued to interact within the scene while it was invisible. We chose these two tasks based on the differing degrees to which they shared common elements with the bowling balls task. The bouncing discs task required tracking real-time dynamics and resolving collision outcomes just as the bowling balls task did. The toppling towers task, on the other hand, required forward prediction from a static snapshot of a single moment (the videos showed the configuration of the towers from all sides, but did not show future physical dynamics playing out). The toppling towers task also required an assessment of the collective behavior of a relatively large set of objects rather than individually tracking just one or two. Despite these differences, variants of the toppling towers task and the bouncing discs task have been shown to engage a common set of brain regions (Fischer et al., 2016), suggesting that they draw on a common "physics engine in the brain."

The toppling towers task and the bouncing discs task each captured reliable individual differences in performance (split-half correlations: $z = 0.44$, $r = .41$ for toppling towers; $z = 0.57$, $r = .52$ for bouncing discs; both $p < .001$), confirming their suitability for use in an analysis of between-task correlations to assess their relationship with mass inference biases. Data from the bowling balls task replicated the findings of Experiment 1, with reliable individual differences in both the bias ($z = 1.22$, $r = .84$, $p < .001$) and discrimination (d-prime; $z = 1.00$, $r = .76$, $p < .001$). Figure 5 shows the pairwise correlations among the four measures (toppling towers, bouncing discs, bowling balls bias, and bowling balls d-prime). The primary comparisons of interest appear along the bottom row of the grid in Figure 5, which shows the relationship between mass inference bias and performance in each of the other three measures. Bias estimates from the bowling balls task were not correlated with any of the other three measures of intuitive physics performance (the smallest p-value was $p = .52$ for the correlation between mass inference bias and performance on the bouncing discs task). Overall accuracy on the bowling balls task (% correct across all trials) was correlated with the measures of bias ($z = 0.27$, $r = .27$, $p = .014$) and d-prime ($z = 1.46$, $r = .90$, $p < .001$), but it confounds the two; for this reason, we focused on the separate measures of bias and d-prime for the remainder of our analyses.

The lack of correlation between mass inference bias and any other measure of intuitive physics performance suggests that the bias

Figure 3

Reliable Individual Differences in Mass Inference Bias and Discrimination

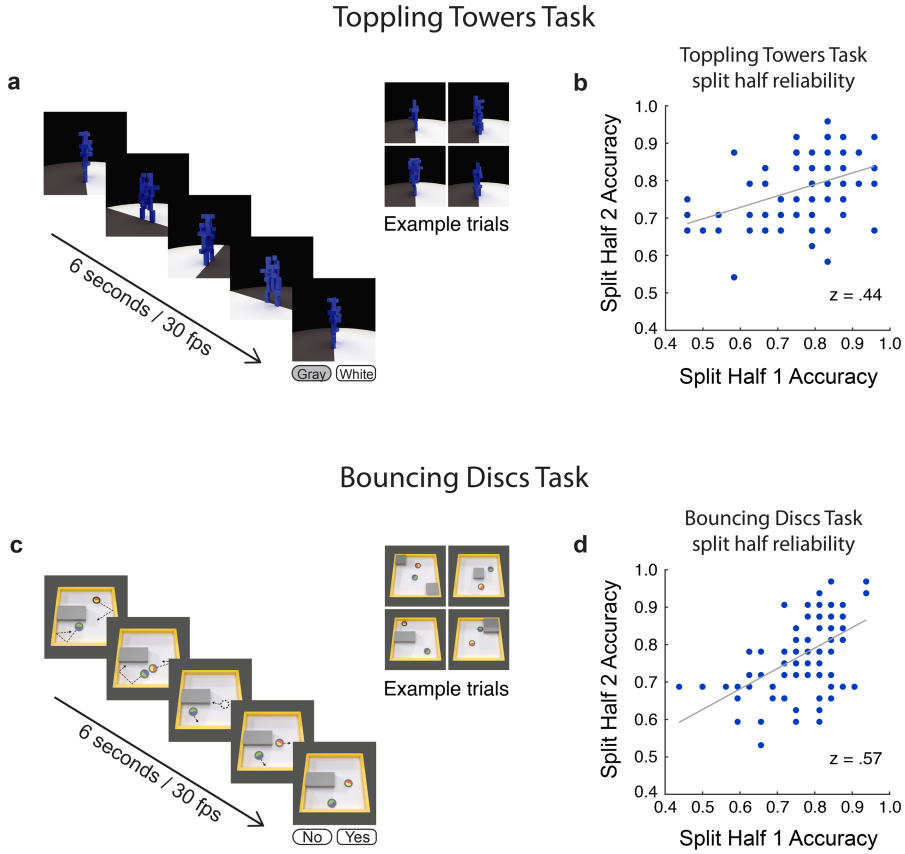


Note. (a) Psychometric curves fit to each individual subject's responses for the full videos. An overall bias is evident from the shift in the point of subjective equality for most individual curves, and substantial individual differences can be seen in the spread of the distribution of curves; (b) shows the split-half reliability of the individual differences in bias, obtained by fitting separate curves to independent halves of each subject's data. The individual differences were highly reliable across independent splits ($z = 0.81$, $r = .67$, $p < .001$; four outliers were excluded from the split-half bias calculation because they fell > 3 standard deviations outside the mean of the distribution). Individual differences in discrimination estimates (d-prime; c) were also highly reliable across split halves ($z = 1.33$, $r = .87$, $p < .001$). (d) Individual differences in both bias and d-prime were also significantly correlated between the full videos and the silhouettes (both $p < .001$), suggesting that a common mental process supports performance on both versions of the task. See the online article for the color version of this figure.

arises from a different source than the fundamental mental operations that support physical reasoning more generally. While this finding by itself is a null result, the case is strengthened by two additional complementary findings. First, as noted above, individual differences in bias estimates were highly reliable, meaning that the lack of correlation with other tasks was not due to a lack of reliable variance. Second, and more importantly, mass discrimination performance (d-prime) measured within the same task as the bias *did* correlate positively and significantly with performance on the bouncing discs ($z = 0.54$, $r = .49$, $p < .001$). A permutation test established that this correlation between mass discrimination and bouncing discs performance was significantly larger than the correlation between mass inference bias and bouncing discs performance ($p < .001$). The correlation between mass inference d-prime and the toppling towers task was not significant ($z = 0.19$, $r = .19$, $p = .077$; discussed below), so our remaining analyses focused on the relationship between the bowling balls and bouncing discs tasks.

As noted above, we included one additional task in Experiment 2—the mental rotations test—as a means of controlling for other possible sources of individual differences not directly related to intuitive physics abilities. Our prior work has shown that performance on intuitive physics tasks is correlated with spatial abilities as measured by classic tasks including the mental rotations test (Mitko & Fischer, 2020). That finding accords with the fact that nearly all scenarios in which we reason about physical behaviors draw on spatial cognition to some degree—objects' geometry, spatial relationships, and movement through space factor critically into physical outcomes. Importantly though, the same study found that intuitive physics performance

Figure 4
Intuitive Physics Tasks Presented in Experiment 2



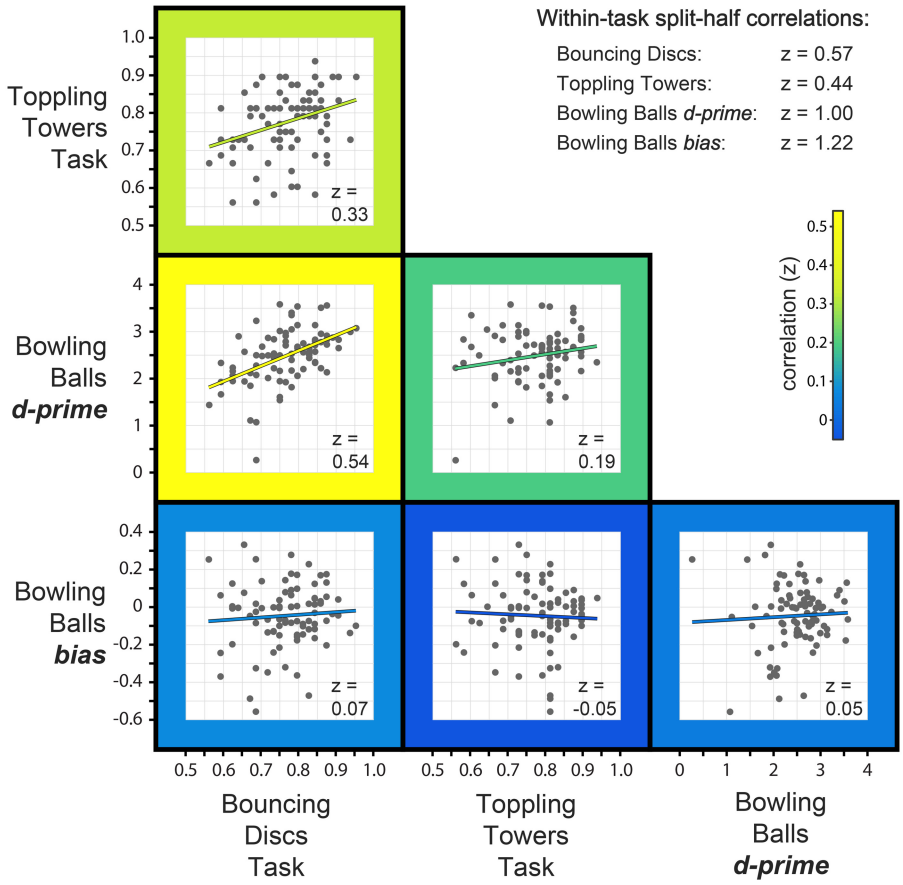
Note. (a) An example trial from the toppling towers task. Each trial presented a 360° camera pan of an unstable block tower that was ready to fall. After viewing each video, participants indicated whether the majority of the blocks would come to rest on the gray or white side of the platform when the tower fell. Inset panels show examples of other block tower configurations. (b) Split-half reliability for performance on the toppling towers task: individual differences in accuracy were significantly correlated across independent split halves ($z = 0.44$, $r = .41$, $p < .001$). (c) An example trial from the bouncing discs task. Each trial showed a movie of discs bouncing within an arena, and participants tracked the anticipated path of one of the discs for a period of 2 s during which it was invisible. Participants responded by indicating whether the disc they tracked had the correct position and velocity when it reappeared. Inset panels show examples of other arena configurations. (d) Split-half reliability for performance on the bouncing discs task: individual differences in accuracy were significantly correlated across independent split halves ($z = 0.57$, $r = .52$, $p < .001$). See the online article for the color version of this figure.

was separable from spatial cognition and varied reliably across individuals after taking individual differences in spatial ability into account. In the present experiment, to be sure that the relationship between mass discrimination and performance on the bouncing discs task was not solely due to the contribution of spatial abilities (or other potential contributors such as effort devoted to the tasks), we conducted a regression analysis in which we factored out individual differences in performance on the mental rotations test. We first evaluated the split-half correlation of performance on the mental rotations test and found reliable individual differences ($z = 1.28$, $r = .86$, $p < .001$ for the split-half correlation). We then regressed out mental rotation scores from both the bowling balls d-prime estimates and bouncing discs performance, obtaining residuals for each task reflecting the variance not explained by performance on the mental rotations test. The residuals for the two tasks were significantly correlated ($z = 0.42$, $r = .40$, $p < .001$), indicating that the relationship between the two was not simply due to individual differences in effort or spatial abilities.

Taken together, these findings suggest that the mass inference bias has a distinct source from the mental operations that support intuitive physics performance more generally, at least with regard to mass discrimination performance in the bowling balls task and prediction performance in the bouncing discs task. We did not find a significant relationship between mass discrimination and performance on the toppling towers task, which would have further bolstered the notion that all intuitive physics tasks here draw on a common resource, distinct from the source of the bias in mass inference. As alluded to above, this result might be expected based on the ways in which

Figure 5

Mass Inference Bias Is Not Related to Intuitive Physics Performance More Generally



Note. Scatterplots show the relationship of individual differences in performance between pairs of tasks in Experiment 2; cells are color-coded according to the strength of each pairwise correlation. Each intuitive physics task (bowling balls, topping towers, and bouncing discs) captured reliable individual differences in performance as indicated by within-task split-half correlations, but an individual's degree of mass inference bias was not predictive of their performance on any task, including mass discrimination (d-prime) in the bowling balls task (smallest p-value = .52). Mass discrimination performance, on the other hand, was significantly correlated with performance on the bouncing discs task ($z = 0.54$, $r = .49$, $p < .001$), and significantly different from the correlation between mass inference bias and bouncing discs (permutation test $p < .001$). These data suggest that the source of the mass inference bias is distinct from the factors that drive intuitive physics performance more generally. See the online article for the color version of this figure.

the toppling towers task differed from the others. To speculate, the role of collisions in the tasks might be particularly important. Although the bowling balls task and the bouncing discs task required different kinds of inference (inferring latent physical properties vs. predicting physical dynamics), both hinged critically on online processing of individual collision events and an understanding of the behaviors immediately following the collisions. In the toppling towers task, on the other hand, some blocks would collide if the towers were allowed to fall, but those collisions would be too numerous and uncertain to interrogate individually and track the precise postcollision behaviors. A comparison of individuals' performance on the bouncing discs and toppling towers tasks further supported the notion that the toppling towers task is somewhat distinct from the others—individual differences were significantly correlated across toppling towers and bouncing discs ($z = 0.33$, $r = .32$, $p = .003$), but the relationship was significantly weaker than the one between mass discrimination and bouncing discs (permutation test; $p = .043$). If the intuitive physics system in the mind comprises a collection of mental operations applied to various kinds of events, the central role of collision analysis in the bowling balls and bouncing discs tasks might be the component that leads to their close relationship in performance.

Experiment 3

If the mass inference bias is not reflective of overall intuitive physics performance, where does it come from? The results of Experiments 1 and 2 are consistent with an account put forward by Sanborn and colleagues (Sanborn, 2014; Sanborn et al., 2013) whereby the bias is a consequence of noisy perceptual estimates.

This prior work showed that if a Newtonian mental model combines noisy velocity estimates with priors about the physical structure of the scene, the resulting mass inferences from observed collisions closely match human judgments, including the bias to overestimate the incoming object's mass. This logic applies equally well to our present experiments, and might explain why the bias measured in Experiment 1 was significantly attenuated for the full videos as compared with the silhouettes—the additional visual cues in the full videos such as the surface texture of the bowling balls might allow for more precise velocity estimates. However, even with the rich visual cues in the full videos in Experiments 1 and 2, participants still displayed reliable biases in their mass inferences. Our findings so far leave open the possibility that the remaining biases have a different source than the bias introduced by noisy perceptual estimates for simplified displays. In Experiment 3, we tested whether the bias observed for the full videos exhibits characteristics of heuristic processing, which would suggest that heuristics and perceptual noise both contribute to mass inference biases when viewing real-world scenes.

The heuristics that have been put forward to account for the bias in mass inference rely largely on a comparison of postcollision velocities; under a heuristic account, performance would be markedly degraded when no such comparison is possible (Sanborn, 2014). On the other hand, if observers employ a Newtonian mental model of physical behavior that incorporates prior beliefs about the balls' velocities, masses, and coefficients of restitution, they should still be able to fall back on these priors when postcollision dynamics are not observable, leading them to perform above chance at discriminating the balls' relative masses *and* display a bias toward perceiving the incoming ball as heavier (Sanborn, 2014). Thus, there is a qualitative difference between the predictions made by the two accounts, with the heuristic account predicting that restricting the visibility of postcollision behaviors will result in a systematic pattern of degraded performance (Figure 6c, inset), while the noisy Newtonian account predicts that performance in the occluded conditions will be similar to the case where all dynamics are visible, if perhaps somewhat less precise overall. In Experiment 3, we tested these predictions using modified versions of the full collision videos in which either the incoming ball or the stationary ball was removed from the video immediately after the collision, preventing a postcollision comparison of velocities (see "Material and Method").

Mass inference performance under three conditions (incoming ball occluded, stationary ball occluded, and no occlusion) is shown in Figure 6, and reveals that participants were still able to discriminate the ball's relative masses under both occlusion conditions, and they still display the same bias toward overestimating the incoming ball's mass. When the incoming ball was hidden postcollision, group accuracy was $M = 0.66$, $SD = 0.21$, with a d-prime of 1.06 (bootstrapped 95% confidence intervals = [0.81, 1.29], $p < .0001$). When the stationary ball was hidden postcollision, group accuracy was $M = 0.71$, $SD = 0.20$, d-prime = 1.42 (bootstrapped 95% confidence intervals = [1.19, 1.66], $p < .0001$). The "no occlusion" condition provides a baseline, showing performance for the full videos when displayed in grayscale to match the occlusion conditions. Overall accuracy for the "no occlusion" videos was 0.72, $SD = 0.21$, d-prime = 1.52 (bootstrapped 95% confidence intervals = [1.25, 1.77], $p < .0001$). We fit logistic curves to the group data from all three conditions. When the incoming ball was hidden, the PSE was −0.11 (bootstrapped 95% confidence intervals = [−0.19, −0.028], $p = .0078$). When the stationary ball was hidden, the PSE was −0.14 (bootstrapped 95% confidence intervals = [−0.22, −0.081], $p < .0001$). In the unaltered condition, the PSE was −0.18 (bootstrapped 95% confidence intervals = [−0.24, −0.13], $p < .0001$). Ten thousand sample permutation tests found that the difference between the hidden incoming ball condition and the unaltered condition was not statistically significant (observed difference = −0.075, 95% confidence interval = [−0.081, 0.080], $p = .067$). Similarly, the difference in PSEs between the stationary ball occlusion condition and the unaltered condition was also not statistically significant (observed difference = −0.040, 95% confidence interval = [−0.071, 0.070], $p = .27$).

Taken together, the results of Experiment 3 show that the mass inference bias we observe when presenting videos of real-world collisions is still consistent with a noisy Newtonian account (and not consistent with heuristics that rely on postcollision velocity comparisons), even if the presence of richer visual cues attenuates the bias somewhat by allowing for more precise perceptual estimates.

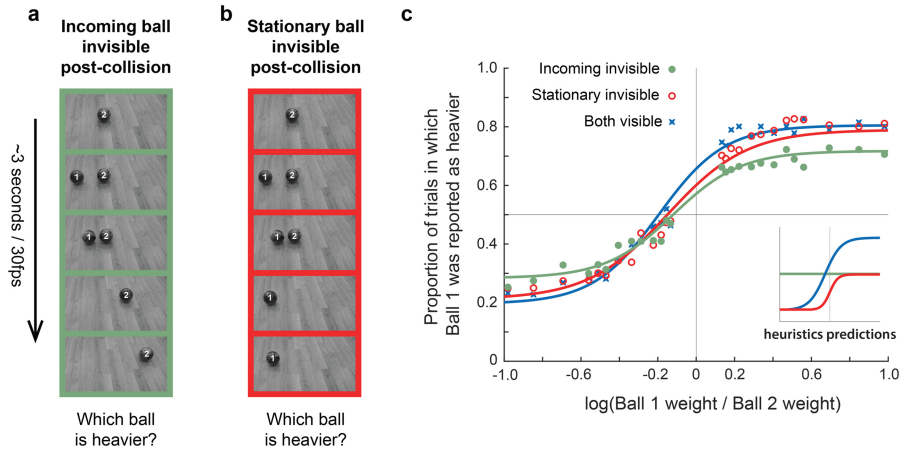
Discussion

The findings from our three experiments provide a window into the source of systematic mass inference errors that have been the subject of scrutiny for decades. These errors, although reliable and variable across individuals, do not arise from the same mechanisms that drive individual differences in intuitive physics performance in general—including individual differences in the sensitivity to relative mass in the same task. Rather, the errors arise from a Newtonian mental model operating on imperfect information. There are many stages at which cognitive and perceptual bottlenecks could limit the information available to the intuitive physics system, and the limitation could go all the way back to the stimulus itself. However, the results of Experiment 1 show that oversimplified stimuli are not a major factor in the emergence of mass inference biases. Rather, the biases can be accounted for by assuming some noise in perceptual estimates and a system that uses prior beliefs to cope with that noise (Sanborn, 2014; Sanborn et al., 2013). The results of Experiment 3 here bear out the predictions of such a model, providing a straightforward account of why people are biased in their mass inferences even though they make predictions that are in line with Newtonian mechanics in a range of scenarios (Bates et al., 2019; J. Hamrick et al., 2011; Ullman et al., 2017).

A crucial component of the current study was the use of videos of real-world collisions. Previous work has shown that using naturalistic stimuli can reveal better performance on intuitive physics tasks where observers typically make systematic errors when viewing simplified stimuli (Kaiser et al., 1986; K. Smith et al., 2013—but see McCloskey & Kohl, 1983 for a counterexample). Likewise, Flynn (1994) made the case that biases in mass inference arose when participants viewed simplified animations of collisions, but not when they viewed movies of real collisions. These findings motivated our decision to use movies of real bowling ball collisions in our experiments here; however, unlike Flynn, we nonetheless observed a robust bias toward overestimating the incoming ball's weight. Why do our findings differ from Flynn's? One possibility that was suggested in follow-up work by Flynn (1995), is that the bias can be speed-dependent, emerging only within a particular range of initial speeds. We did not experimentally manipulate the starting speed in our experiments (and we verified that the incoming speed was

Figure 6

Experiment 3: Removing Visual Cues Critical for Heuristic-Based Mass Inferences Leaves Performance Intact



Note. (a) Videos with the incoming ball hidden from view postcollision. Using grayscale versions of the full videos from Experiments 1 and 2, we identified the frame of each video in which the collision occurred and digitally removed the incoming ball (Ball 1) from the collision frame and every frame thereafter. This manipulation prevented participants from seeing the ricochet of Ball 1 and from comparing the velocities of the balls after the collision occurred. (b) Videos with the stationary ball hidden from view postcollision. In a second set of videos, we used the same process as above to hide the initially stationary ball from view from the time of collision onward. This manipulation prevented participants from seeing any motion of Ball 2, rendering a postcollision velocity comparison impossible. (c) Psychometric curve fits to the group data from the videos with the incoming ball removed from view (green filled circles), the videos with the stationary ball removed from view (red open circles), and videos in which both balls remained visible for the full duration of the events (blue xs). The inset panel shows qualitative predictions for performance under the three conditions if observers rely on heuristics based on postcollision dynamics (adapted from Sanborn, 2014). While the heuristics could support mass discrimination comparable to what we observed in Experiments 1 and 2 when both balls remain visible (blue prediction curve), solely relying on the heuristics would lead to degraded performance when either ball is occluded postcollision. When the incoming ball is rendered invisible, it is not possible to compare postcollision velocities or observe whether the incoming ball ricochets, so the heuristics provide no information about the balls' relative masses (green prediction curve). When the stationary ball is occluded postcollision, only the ricochet heuristic provides useful information, leading to chance performance across part of the stimulus range (red prediction curve). In contrast to the heuristics predictions shown, participants' performance in both occlusion conditions was comparable to the case of no occlusion, and in all conditions, participants displayed a bias toward overestimating the mass of the incoming ball. See the online article for the color version of this figure.

very similar across all videos and not correlated with mass), so it is possible that our movies had initial speeds that were more likely to reveal observers' biases than the videos in Flynn's initial work (Flynn, 1994). Critically though, the existence of a systematic bias for *any* range of initial speeds is evidence against the notion that the bias emerges from a lack of real-world cues, whereas in a noisy Newtonian framework, the strength of the bias emerging from a Newtonian model operating on noisy inputs would in fact be dependent on the initial speed of the incoming ball.

Our findings from Experiment 2 suggest that mass inference draws on general intuitive physics mechanisms that also support other kinds of inferences and predictions. Even though the nature of the inferences is different, they rely on common Newtonian principles and can draw on a common set of algorithms. More work is needed to establish just how flexible these general intuitive physics resources are, and whether there are particular physical scenarios (e.g., the behaviors of fluids and soft bodies) that we treat as special cases and reason about separately from rigid body dynamics. Our results join other previous findings of a connection between the mental processes for inferring mass and predicting dynamics. J. B. Hamrick et al. (2016) used a version of the towers task to assess how mass inference and stability predictions are intertwined. Subjects were able to accurately infer which blocks in a tower were heavier by observing the stability (or instability) of the tower. J. B. Hamrick et al. (2016) also found a systematic association between these mass inferences and earlier predictions of tower stability, in line with the current study's evidence for a shared mechanism needed to accomplish both types of physical reasoning.

Experiment 3 here used an established approach (Sanborn, 2014) to evaluate the tenability of a heuristic account of mass inference that relies on a comparison of postcollision speeds and the existence (or not) of a ricochet after the collision. We evaluated these particular heuristics because they are the first and most

widely-cited heuristic account of performance on the task (Gilden & Proffitt, 1989, 1994). There are other rules that could be applied to similar effect, and exhaustively ruling out all possible rule-based accounts would be an unending endeavor. We would simply stress that a heuristic account is not a priori more parsimonious than a robust Newtonian one, despite the simplicity of individual rules that might compose a library of heuristics. A Newtonian account has the appeal of being able to accommodate a broad variety of scenarios with a compact set of flexible principles. There are particular scenarios in which we very likely rely on heuristics (e.g., closed containers, Davis et al., 2017; Davis & Marcus, 2016), and these cases come with well-specified rules ("nothing leaves a sealed container") that are correct in the vast majority of cases we encounter. We propose that people revert to heuristic physical reasoning in lieu of forming detailed Newtonian predictions in the following scenarios: (a) when scene content (i.e., the geometry and material properties of the relevant objects and surfaces) cannot be represented with sufficient fidelity to perform a process like simulation. Examples would be scenes with convoluted 3D structure, large numbers of objects that surpass the capacity of working memory, or a large number of relevant spatial locations that cannot be simultaneously attended. In these cases, bottlenecks in perception and cognition constrain the information available to the intuitive physics system and necessitate heuristic strategies for physical reasoning. (b) Cases where a simple heuristic is essentially always correct. For example, an item sealed in a container will remain inside. (c) Inferences that are overlearned through extended training. For example, the fast, highly precise physical judgments made by athletes, craftspeople, and others in the course of their work. Building expertise in these fields may entail developing a library of rules that allow for bypassing general intuitive physics mechanisms. These scenarios are common in daily life, as are scenarios that are well-suited to general intuitive physics mechanisms. We may continually employ a hybrid of heuristic reasoning and Newtonian inference to understand and predict the physical behaviors of our everyday environments.

A potential constraint on the generality of our findings comes from the approach of using computerized tasks rather than real-world scenarios viewed in person. The ultimate goal of this work is to understand how people infer the physical structure and dynamics of their everyday environments, and scenarios presented onscreen differ from those encountered in person even when the stimuli consist of videos of real-world events as in the present study. One key difference is that real-world physical events often carry a host of action affordances. For example, if you saw a collision between bowling balls in real life (most likely during an outing at the bowling alley), you might reach to catch one of the balls as it rolls down the ball return or extend a foot to stop a stray ball on the floor from rolling away. There is evidence that engaging in these kinds of natural actions can affect the accuracy and precision of physical inferences (Schwartz & Black, 1999; K. Smith et al., 2013). Understanding when and why action generation interacts with physical inferences will be a key piece of the puzzle in understanding intuitive physics more broadly (Fischer & Mahon, 2021), and it will be important for future work to extend our present tasks and others into the realm of real-world interactive experiments. For the experiments presented in this article, we opted for computerized presentation because it afforded some practical advantages over presenting real-world, in-person physical events: we were able to test a much larger sample of participants than would have been feasible with a real-world setup that had to be reset after each trial, and we were able to repeat the events with perfect reliability so that the available information was consistent across trials and participants.

In sum, our findings here show that although people make systematic errors when judging the masses of interacting objects, the errors are not evidence of a broken system. The source of the errors is distinct from the general intuitive physics mechanisms that support physical inferences in a variety of tasks, and likely arises from uncertainty in perceptual estimates of the scene attributes that serve as inputs to the intuitive physics system. These findings reconcile the apparent discrepancy between the existence of such errors and a wave of new findings showing that people's intuitions about physicality often fall in line with Newtonian mechanics.

Material and Method

Experiment 1

Participants

For all experiments, the participants were recruited on Amazon's Mechanical Turk (mTurk). Participants were required to be 18–35 years of age and complete the study from within the United States. All participants provided anonymously informed consent in accordance with Johns Hopkins University IRB protocols. In Experiment 1, the full videos and silhouette videos were tested within subjects in a blocked design. A total of 153 Mechanical Turk workers completed the tasks, but a total of 67 of them were excluded before analyses based on preplanned exclusion criteria. One participant did not complete the task from within the United States, 12 participants had missing trials due to internet connection problems, and eight had completed an earlier version of the task (the task description instructed online workers not to complete the current task if they had participated in one of the listed previous studies). We also applied a performance-based cutoff with the aim of excluding participants who ignored or did not understand the task instructions. In prior piloting, we found that performance for the largest weight differences was high (typically 85%–100% correct) in all participants who faithfully performed the task. We used these trials with the largest weight differences (6 lb incoming ball with 14 and 16 lb stationary ball and 16 lb incoming ball with 6 and 8 lb stationary ball) as "catch" trials, and required that a participant performs at 55% correct or better on these trials to be included in the subsequent analyses. We used a lenient accuracy cutoff to selectively exclude participants who did not perform the task as instructed while allowing participants who attempted the task—even if they had difficulty with it—to remain in the analysis. Forty-six participants were excluded based on this performance cutoff. Finally, one participant was excluded for giving the same response on >90% of trials. After exclusions, there were a total of 85 participants for analysis. Forty-five of these participants were shown a block of full videos first, and the other 40 participants were shown a silhouette block first. Participants were paid $8 to complete the mTurk HIT.

Stimuli

To create the collision videos, we used six "Brunswick Tzone Deep Space" bowling balls in indigo color. The weights of the

balls were 6, 8, 10, 12, 14, and 16 pounds. Note that although all the balls were of the same brand and color, there was natural variation in the "wavy" coloring of each ball. We opted to use bowling balls with patterning in their surface coloration to provide a visual cue to the rotation of the balls in the stimulus videos.

In each collision video, one ball was placed in a marked position on the floor (stationary ball). The other ball (incoming ball) was held by an experimenter at the top of a wooden ramp outside of the camera view. When the incoming ball was let go, it rolled down the ramp and collided with the stationary ball. This procedure was completed twice for all 30 possible pairs of weights, totaling 60 stimuli. All collisions were filmed in 30 frames/s and the sound of the videos was removed. Each video ended shortly after one of the balls left the frame. Each video lasted approximately 3 s.

To create the silhouette stimuli, the position and size of each ball were annotated in Matlab frame-by-frame. The videos were recreated in Matlab, using dark gray circles/ellipses for the bowling balls and removing the background image sans a few lines on the floor for perspective.

Design

In a blocked design, participants saw each stimulus of the full videos and the silhouette version 4 times, totaling 480 trials. In each block, participants would see 60 trials of either the full videos or the silhouette videos. After each block, participants could take a self-timed break before continuing to the next block. The type of block alternated back and forth with half of the participants seeing full videos first and the other half seeing silhouette videos first.

On every trial, participants would see a collision video in the middle of the screen. Participants would then have the opportunity to select a button labeled "BALL 1" or "BALL 2," indicating which ball was heavier. In the task instructions, participants were informed that ball 1 always referred to the incoming ball and ball 2 always referred to the stationary ball. The question "Which ball is heavier" remained on the screen at all times to remind participants of the task judgment. After making a selection, participants pressed another button to advance to the next trial. This allowed the participants to potentially change their answers, though they could not view the video again.

Analysis

Participants' accuracy for performance-based exclusion criterion was calculated for the full video trials and the silhouette videos separately, using only the trials with the largest weight difference between the incoming and stationary bowling balls (6 lb incoming ball with 14 and 16 lb stationary ball and 16 lb incoming ball with 6 and 8 lb stationary ball).

As a measure of mass discrimination performance, we calculated a d-prime for each participant using a log-linear approach (Hautus, 1995). To compute each participant's bias in mass inference, we fit a psychometric curve to each individual participant's data. We first binned the trials by the weight ratio of the incoming ball to the stationary ball and computed the proportion of the time that the participant reported the incoming ball as heavier within each bin. We then fit a logistic curve of the form:

$$\text{Proportion "incoming heavier" response} = \frac{1-2c}{1+e^{-a(\text{weightratio}-b)}} + c$$

The a parameter in the curve fit scales the steepness of the curve, the b parameter gives the bias, and the c parameter gives the lapse rate (upper and lower asymptotes). Separate curves were fit to the full video data and the silhouette video data.

To compute confidence intervals and test the significance of the bias in mass inference, we used a combination of bootstrapping and permutation testing. On each of the 10,000 bootstrapped iterations, logistic curves were fitted as described above and the bias (b parameter) was recorded. We tested whether the resulting bootstrapped distribution significantly differed from zero (two-tailed test) to test the significance of the bias in a single condition, and we used a permuted null distribution to test for a significant difference in bias between two conditions. To generate a permuted null distribution, on each of the 10,000 iterations, we permuted the condition labeling while maintaining the same number of trials in each weight ratio bin for each labeled condition. We then fit a logistic curve to the data for each (permuted) condition and recorded the difference in bias estimates between the two conditions. The true (unpermuted) difference in bias estimates was tested against this null distribution (two-tailed test) to test whether the observed difference in bias estimates between the two conditions differed significantly from that which would be expected by chance (Ernst, 2004).

To assess test–retest reliability, we performed split-half analyses. In each of the 10,000 iterations, the trials for each participant were randomly split in half, maintaining the same distribution of trial types in each half of the data. We then computed the measure of interest independently in each half. We plotted all participants' split-half data together (first half vs. second half, which each point representing one participant), and computed the correlation between halves as a measure of reliability. In all analyses in which correlations were computed, the resulting r-values were transformed to Fisher z-values using the transform $z = 0.5 \times \log((1+r)/(1-r))$. We express correlations as z-values rather than r so that they may be linearly compared in magnitude—z is unbounded, whereas r values are bounded in the range of (0 1).

Experiment 2

Participants

In Experiment 2, each participant completed four tasks: the bowling balls weight inference task as in Experiment 1, the toppling towers task, the bouncing discs task, and the mental rotations test. The order of the four tasks was randomized for each participant, and participants were required to complete one task at a time. In total, 183 mTurk workers completed the tasks. Fifty-five participants were excluded before analyses for the following reasons: 16 participants had malfunctions with the data saving file, three participants did not complete the task from within the United States, 14 participants had completed an earlier version of one or more of the tasks, and 22 participants attempted to complete multiple tasks at the same time. Additionally, 42 participants were excluded for performing below 55% on one or more of the tasks. After exclusions, a total

of 86 participants remained for analysis. Participants were paid $9.50 to complete the mTurk HIT.

Tasks

Bowling Balls Mass Inference. The Bowling Balls task in Experiment 2 was identical to the task in Experiment 1 using the full bowling ball collision videos. Participants saw all 60 stimuli 4 times each for a total of 240 trials. After each video, participants judged which bowling ball was heavier.

Toppling Towers Task. The toppling towers task (Mitko & Fischer, 2020) is a modified version of the unstable towers task (Battaglia et al., 2013; Fischer et al., 2016; J. B. Hamrick et al., 2016) in which subjects must predict how an unstable tower of blocks will tumble. In every trial, participants viewed an unstable tower in a 360° panoramic video that lasted for 6 s at 30 frames/s (Figure 4a). The towers were centered on a platform that was split in the middle by color (either gray or white). Participants had to judge on which side the majority of the blocks would come to rest after the tower had fallen. Participants first saw a practice trial that showed a full video of a tower falling so that they could develop a sense of the materials and mass used for the blocks. They then viewed 48 tower videos and made judgments about how the blocks would fall without receiving feedback. After viewing each video, participants made their responses by clicking buttons under the two sides of the platform labeled "Gray" or "White." Participants were required to make a response in order to advance to the next trial. The number of blocks used to construct the towers varied: there could be 11, 13, 15, 17, 19, or 21 blocks within a tower. There were eight towers for each block number, with four falling to the gray side and four falling to the white side. We presented the towers in a random order to each participant. The last frame of each video remained on the screen until the participant advanced to the next trial. For more information on the toppling towers task, see Mitko and Fischer (2020).

Bouncing Discs Task. The bouncing disc task is a new task that requires participants to track the velocity of a disc after it has disappeared. On every trial, participants viewed two discs moving in a square "arena." After 2 s, one of the discs would disappear from the scene. After 2 more seconds, the disc would reappear, and the video would play for a final 2 s. Participants then selected either "Yes" or "No" to answer whether the disc reappeared with the correct position and velocity. Stimuli were made using Blender 3-D modeling software (http://www.blender.org), and physical outcomes were assessed using simulations run in Blender's built-in bullet physics engine. Four different arenas were used, each being square but with varying internal barriers. There were 16 different stimuli made in each arena for a total of 64 stimuli, each with a randomized starting location and velocity. The stimuli were counterbalanced for which disc would disappear and whether it would reappear correctly. In the incorrect reappearance stimuli, the reappearing disc would reappear in a location near the correct location, with a randomly altered velocity. During pilot studies, we identified stimuli in which performance ranged from 55% to 95%, and only used these stimuli for the final version of the task. Participants performed three practice trials that gave feedback by showing how the stimulus would look without one of the discs disappearing.

Mental Rotations Test. To control for other factors that may have influenced individual differences in the physics tasks, such as overall effort and spatial abilities, subjects also performed the mental rotations test (Peters et al., 1995; Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). Subjects were shown 24 stimuli of connected blocks and were asked in each question which two options out of a possible four were rotated versions of the target stimulus. The two incorrect options were structurally different objects. Subjects were shown correctly answered questions prior to starting the test. There was no time limit for completing it and subjects could change their responses at any time. Accuracy was calculated by dividing the total number of correct choices by 48, as there were two correct answers for each question.

Experiment 3

Participants

Experiment 3 included three conditions in a between-subjects design: (3a) incoming ball occlusion, (3b) stationary ball occlusion, and (3c) no occlusion. Using the same preanalysis exclusion criteria as in Experiment 1, we excluded 31 participants before the analysis. The breakdown of the excluded participants was as follows: four participants did not complete the task from within the United States, and 27 had completed an earlier version of the task. As in Experiments 1 and 2, we also sought to exclude participants who ignored or did not understand the task instructions. However, unlike the previous experiments where our piloting led us to expect that any participant genuinely attempting the task would perform well on those "catch" trials, we did not make the same assumption in Experiment 3. Experiment 3 evaluated the hypothesis that hiding the balls' postcollision behavior would impair the performance, and in this case, it was possible that even participants who genuinely attempted the task would have performance near or at chance. For this reason, we only excluded participants who gave the same response on >90% of trials, indicating that they ignored the instructions. Six participants were excluded based on this criterion. After all exclusions, there was a total of 91 participants in the incoming ball occlusion condition, 94 participants in the stationary ball occlusion condition, and 88 participants in the no occlusion grayscale condition. Participants were paid $2 to complete the mTurk HIT.

Stimuli

In condition 3a, the incoming ball disappeared from the video postcollision. To create these video stimuli, an empty background frame of the floor masked the motor ball (and its shadow) starting the frame after the collision. The same procedure was done for condition 3b, except the stationary ball was masked following the collision. All of the videos were shown in grayscale as the experimenters found this hid lighting artifacts of the disappeared ball. For an authentic control comparison, condition 3c showed grayscale versions of the unaltered videos.

Analysis

The exclusion criteria for Experiment 3 were similar to Experiment 1, except we did not exclude subjects for performance on the "easiest" trials. Logistic curves were fit to the group data for each condition in the same manner as in Experiment 1. To determine whether the bias was significantly different than 0, logistic curves were fit to 10,000 bootstrapped samples for each condition, as described in Experiment 1. To test for differences in the bias

between the motor ball occlusion and the unaltered video conditions, a permutation test was performed with 10,000 samples as described in Experiment 1. The same method was also used to test for a significant difference between the bias in the stationary ball occlusion and the unaltered video conditions.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), Article e1007210. https://doi.org/10.1371/journal.pcbi.1007210

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 of the United States of America*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72. https://doi.org/10.1016/j.artint.2015.12.003

Davis, E., Marcus, G., & Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial Intelligence*, *248*, 46–84. https://doi.org/10.1016/j.artint.2017.03.004

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, *19*(4), 676–685. https://doi.org/10.1214/088342304000000396

Fischer, J. (2020). Naïve physics: Building a mental model of how the world behaves. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The cognitive neurosciences VI* (pp. 777–783). MIT Press.

Fischer, J., & Mahon, B. Z. (2021). What tool representation, intuitive physics, and action have in common: The brain's first-person physics engine. *Cognitive Neuropsychology*, *38*(7–8), 455–467. https://doi.org/10.1080/02643294.2022.2106126

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Flynn, S. B. (1994). The perception of relative mass in physical collisions. *Ecological Psychology*, *6*(3), 185–204. https://doi.org/10.1207/s15326969eco0603_2

Flynn, S. B. (1995). Effect of velocity on judgments of relative mass. *Perceptual and Motor Skills*, *81*(3), 979–987. https://doi.org/10.2466/pms.1995.81.3.979

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372–383. https://doi.org/10.1037/0096-1523.15.2.372

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, *56*(6), 708–720. https://doi.org/10.3758/bf03208364

Guo, L., Courtney, S. M., & Fischer, J. (2020). Knowledge of objects' physical properties implicitly guides attention during visual search. *Journal of Experimental Psychology: General*, *149*(12), 2332–2343. https://doi.org/10.1037/xge0000776

Hamrick, J., Battaglia, P., & Tenenbaum, J. B. (2011). *Internal physics models guide probabilistic judgments about object dynamics*. Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA.

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. https://doi.org/10.1016/j.cognition.2016.08.012

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*(4), 308–312. https://doi.org/10.3758/bf03202508

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. https://doi.org/10.1016/j.tics.2017.06.002

Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology*, *19*(3), 342–368. https://doi.org/10.1016/0010-0285(87)90008-9

McCloskey, M., & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 146–156. https://doi.org/10.1037/0278-7393.9.1.146

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649. https://doi.org/10.1037/0278-7393.9.4.636

Mitko, A., & Fischer, J. (2020). When it all falls down: The relationship between intuitive physics and spatial cognition. *Cognitive Research: Principles and Implications*, *5*(1), Article 24. https://doi.org/10.1186/s41235-020-00224-7

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test–different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58. https://doi.org/10.1006/brcg.1995.1032

Runeson, S., & Frykholm, G. (1981). Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 733–740. https://doi.org/10.1037//0096-1523.7.4.733

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review*, *107*(3), 525–555. https://doi.org/10.1037/0033-295X.107.3.525

Runeson, S., & Vedeler, D. (1993). The indispensability of precollision kinematics in the visual perception of relative mass. *Perception & Psychophysics*, *53*(6), 617–632. https://doi.org/10.3758/BF03211738

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology*, *5*, Article 938. https://doi.org/10.3389/fpsyg.2014.00938

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411–437. https://doi.org/10.1037/a0031912

Schwartz, D. L. (1995). Reasoning about the referent of a picture versus reasoning about the picture as the referent: An effect of visual realism. *Memory & Cognition*, *23*(6), 709–722. https://doi.org/10.3758/BF03200924

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116–136. https://doi.org/10.1037/0278-7393.25.1.116

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Smith, K., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th Conference of the Cognitive Science Society* (pp. 3426–3431). Cognitive Science Society.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199. https://doi.org/10.1111/tops.12009

Todd, J. T., & Warren, W. H., Jr. (1982). Visual perception of relative mass in dynamic events. *Perception*, *11*(3), 325–335. https://doi.org/10.1068/p110325

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*(9), 649–665. https://doi.org/10.1016/j.tics.2017.05.012

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599–604. https://doi.org/10.2466/pms.1978.47.2.599

Vicovaro, M., Noventa, S., & Battaglini, L. (2019). Intuitive physics of gravitational motion as shown by perceptual judgment and prediction-motion tasks. *Acta Psychologica*, *194*, 51–62. https://doi.org/10.1016/j.actpsy.2019.02.001

Zhang, A., Cormiea, S., & Fischer, J. (2019). Weight and see: Vicarious perception of physical properties in an object lifting task. *Journal of Vision*, *19*(10), Article 219a. https://doi.org/10.1167/19.10.219a

Received January 26, 2022
Revision received January 16, 2023
Accepted February 3, 2023 ■