Ensemble Physics: Perceiving the Mass of Groups of Objects is More Than the Sum of Its Parts

Vicente Vivanco[†], **Joshua B. Tenenbaum, Vivian C. Paulun**[*], **& Kevin A. Smith**[*]
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA, USA
[†] vvc@mit.edu

Abstract

Imagine pouring a box of granola into a bowl. Are you considering hundreds of individual chunks or the motion of the group as a whole? Human perceptual limits suggest we cannot be representing the individuals, implying we simulate 'ensembles'of objects. If true, we would need to represent group physical properties beyond individual aggregates, similar to perceiving ensemble properties like color, size, or facial expression. Here we investigate whether people do hold ensemble representations of mass, using tasks in which participants watch a video of a single marble or set of marbles falling onto an elastic cloth and judge the individual or average mass. We find first that people better judge average masses than individual masses, then find evidence that the better ensemble judgments are not just due to aggregating information from individual marbles. Together, this supports the concept of ensemble perception in intuitive physics, extending our understanding of how people represent and simulate sets of objects.

Keywords: ensemble perception; intuitive physics

Introduction

Every morning, many people pour granola from a box into their cereal bowl with little thought or effort. Doing so requires predicting how the granola will travel from the box when it is tipped. But it seems unlikely that we predict the motion of each individual grain or particle of granola, as there are far too many to keep track of given the limits of our attention and working memory. How then are we able to predict the motion of large collections of objects?

While much recent research on intuitive physics has focused on people's predictions about the motions or interactions of a small number of individual objects (see Smith et al. (2024) for an overview), there is evidence that people treat large collections of objects as different from individuals (Fig. 1). For instance, it has been theorized that adults process pouring sand more like a liquid than a collection of solid objects (Kubricht et al., 2017) and even 5-months-old infants show different expectations about the behavior of "stuff", such as piles of sand, compared to solid objects (Hespos, Ferry, Anderson, Hollenbeck, & Rips, 2016). Recent neuroimaging work found that functionally distinct brain regions are engaged when adults observe granular and liquid substances versus solid objects (Paulun, Pramod, & Kanwisher, 2023). However, none of these studies have directly tested whether the mind treats a collection of objects as something distinct and different from a large set of individual items.

Here, we test whether ensembles of objects are represented and processed differently than sets of individual items for physical reasoning. That is, is the representational whole for an ensemble of objects or particles in intuitive physics any different than the sum of its parts? While we are not making claims for precisely *how* these collections might be processed, any theory of ensemble physics would require people to represent properties of the ensemble qualitatively differently from the aggregate of its individual elements: it is not about the position of the constituent items, but the average location and extent of the group as a whole. For example, if asked to predict whether falling sand will be caught by or rip through a sheet of paper, it is not that people represent latent properties like mass for each individual grain, but that they have some notion of the average weight.

There is good reason to believe that people can extract these ensemble properties. Research on *ensemble perception* shows that humans can extract global statistics of orientation, size, and facial expressions without encoding all individual items in a set (Whitney & Yamanashi Leib, 2018; Haberman & Whitney, 2007, 2009; Ariely, 2001). This allows observers to form rapid scene-level inferences and helps conserve attentional resources (Alvarez & Oliva, 2008). However, in contrast to orientation or size, physical properties like mass, elasticity, or softness cannot be extracted from a single image but instead can be inferred from watching objects interact (Paulun, Schmidt, van Assen, & Fleming, 2017; Paulun & Fleming, 2020; Sanborn, Mansinghka, & Griffiths, 2013; Yildirim, Smith, Belledonne, Wu, & Tenenbaum, 2018). If people represent collections of physical objects as *ensembles*, they should be able to use physical reasoning to extract the physical properties of these groups. Here, we test this hypothesis by studying whether people make ensemble judgments of mass.

In particular, we tested whether people can perceive an ensemble's mass in ways that cannot be reduced to aggregating information from individual elements. We present two preregistered experiments that compare single-object mass judgments to ensemble mass judgments. First, we demonstrate that the perception of the average mass of an ensemble is more accurate than the perception of masses of individual objects. We then show that this ensemble benefit is not simply due to aggregating information from individual objects. Together, this shows that people are able to perceive ensemble physical properties, suggesting that the simulation of object sets is different from the simulation of individual objects.

Experiment 1

We first tested whether participants can discriminate mass differences more accurately from ensembles than from individ-

[*]VCP and KAS contributed equally to this work

Figure 1: From playing with marbles, pouring cereal into a bowl, or watching leaves cascade through the air in autumn, humans frequently encounter collections of objects that behave as cohesive groups. Despite the improbability of perceiving and simulating each object individually, people intuitively and accurately predict the behavior of these ensembles. This ability suggests the existence of a perceptual mechanism that extracts ensemble-level properties, enabling judgments about a group's physical characteristics without relying on detailed representations of its individual components.

ual objects. We investigated these discrimination abilities by showing people videos of a single marble or set of 25 marbles falling onto and thus deforming an elastic cloth (see Fig. 2), and asking which of two videos contained the heavier marble(s). Because the deformation of the cloth is based on both the mass of the marble and the speed at which it contacts (which in turn is based on the height at which it is dropped from), watching the full interaction should provide sufficient information to infer the marbles' masses.

We hypothesized that if participants had an understanding of ensemble properties, they would show higher discrimination accuracy when judging the average mass of ensembles of 25 marbles, compared to single marbles, even when the underlying mass differences were matched.

Participants We collected data from 100 participants on Prolific. The experiment lasted approximately 25 minutes, and each participant was compensated $6.25 for their time.

Stimuli All stimuli were created using Houdini, a 3D animation and simulation software. Each stimulus was a 2-second video clip showing marbles falling onto a cloth subdivided into a 5×5 grid. In the simulated scene, the cloth spanned a width of 3.33 m, and marbles were dropped from heights uniformly randomized between 1.3 and 1.6 m. Within each grid square, marble positions were slightly perturbed by up to ± 0.05 m in the x and z directions, and each marble was randomly rotated to avoid texture alignment cues. In the videos, balls would always fall at $9.8 m/s^2$, in line with Earth's gravity. The camera was positioned 6.25 m from the cloth, such that the cloth took up 24.5% of the video window.

There were two types of trials: "Single" trials contained two videos showing only a single marble each, and "Ensemble" trials contained two videos showing 25 marbles each. In both types, one of the two videos was the *reference* video in which the marbles had the same mass (Single) or geometric mean of masses (Ensemble) in all trials. The other video, the *test* video, showed marbles that had a mass or average mass that ranged between $\frac{1}{4}$ to 4 times the reference mass, sampled from 21 possibilities, equally spread in log-mass space. In the Ensemble trials, the 25 marbles had masses such that the geometric mean of all marble masses was chosen in the same way as the Single trials, but the individual marble masses were chosen to produce a log-normal distribution of masses ranging from 0.45 to 2.24 times the average mass; these marbles were distributed randomly across grid positions. We chose to space mass stimuli in log-space because prior psychometric studies have found that people discriminate mass ratios, not absolute mass differences (Sanborn et al., 2013).

In the Single trials, the reference marble was always dropped onto the center of the grid, while the alternate marble's position on the grid was randomized. In all cases, the marbles' initial heights, positions within the grid, and rotations were randomized, requiring participants to rely on mass-related cues rather than extraneous visual information.

Each participant completed 120 trials (60 Single, 60 Ensemble). Trial order and the order of the two videos (within each trial) were randomized.

Procedure Participants began with a practice session consisting of two trials and a brief quiz before the main experiment. The experiment was divided into two blocks, each containing only Single or Ensemble Trials, with block order randomized. On each trial, two sequential videos appeared, separated by a short interval. After watching the two videos in sequence, participants clicked one of two buttons to indicate whether the second stimulus was "Lighter" or "Heavier" relative to the first. The order of the reference and test videos was randomized for each trial. Mass judgments were recorded relative to the reference video.

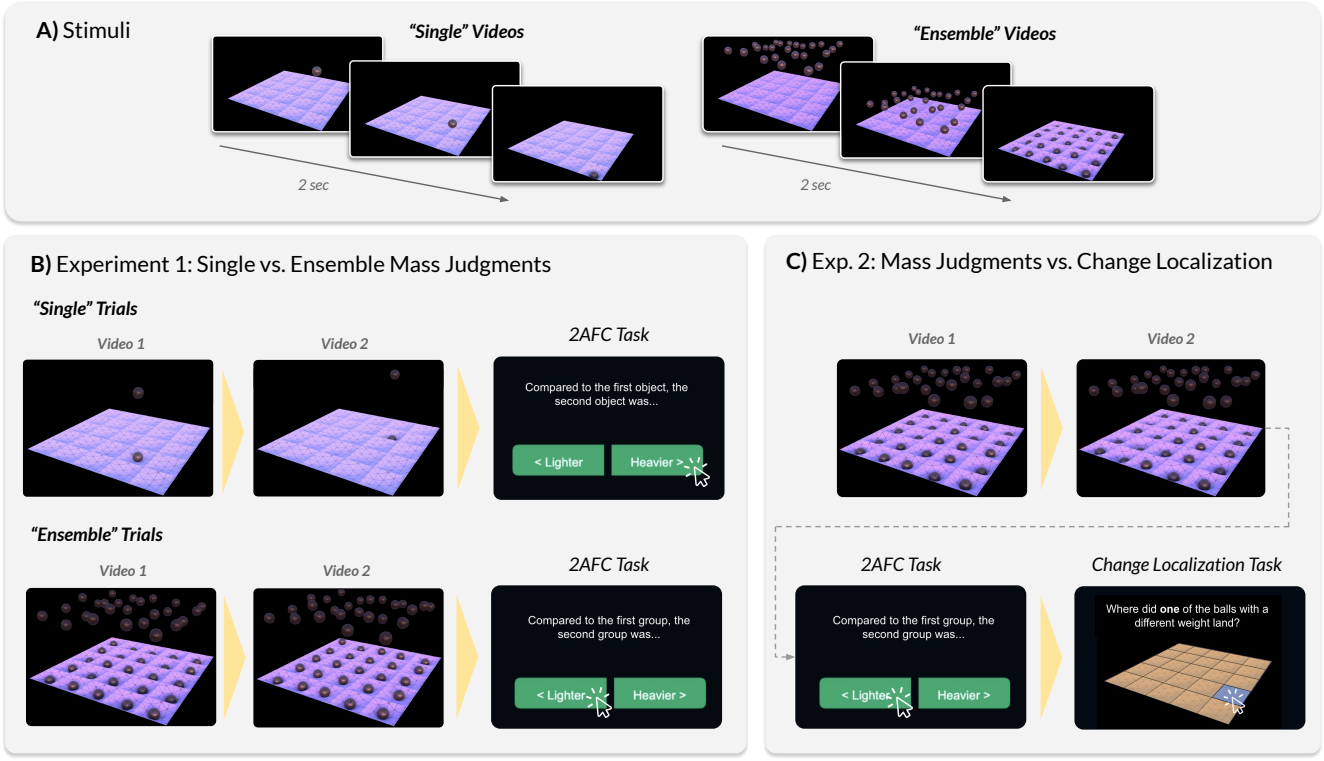

Figure 2: **(A)** Our stimuli consisted of 2s videos showing an individual marble or an ensemble of marbles falling onto a cloth. **(B)** In Experiment 1, participants viewed two sequential videos in each trial and made a two-alternative forced choice (2AFC) indicating which stimulus was heavier. In the "Single" section (top row), each video contained one marble falling onto a cloth. In the "Ensemble" section (bottom row), each video contained 25 marbles. This setup tested whether participants discriminate the mean mass of ensembles more effectively than individuals. **(C)** In Experiment 2, participants saw videos of two ensembles, with the only difference between the two being the mass of 5 marbles. After watching the sequence, participants had to first decide which group was heavier, then pick out one of the five marbles that had different masses between the two videos.

We estimated the shape of the psychometric function for each participant linking the mass ratios between the test and reference videos with the probability that a participant would judge the test video as heavier. The function linking these two variables was a cumulative normal curve with a lapse rate, using separate functions for the Single and Ensemble stimuli.

We used QUEST+ (Watson, 2017), an adaptive Bayesian method, to efficiently select the most informative stimuli to fit the slope, threshold, and lapse rate for each participant. Thus, no two participants saw exactly the same stimuli.

Results We conducted a hierarchical Bayesian analysis to model participants' responses. Each trial's response (y_{ij}) was modeled using a cumulative Gaussian psychometric function:

$$P(y_{ij}=1) = \frac{\lambda_i}{2} + (1-\lambda_i)\cdot\Phi\left(\frac{\log x_{ij} - \mu_i}{\sigma_i}\right) \qquad (1)$$

where $y_{ij}=1$ indicates a "Heavier" response, x_{ij} is the presented mass ratio on trial j for participant i, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and λ_i, μ_i, and σ_i represent the lapse rate, threshold, and slope parameters, respectively.

The hierarchical model included three key participant-specific parameters:

- **Threshold** (μ_i): The log mass ratio at which participants responded "Heavier" 50% of the time, representing the point of subjective equality (PSE).
- **Slope** (σ_i): The standard deviation of the cumulative Gaussian. Smaller σ_i indicates higher sensitivity to mass differences.
- **Lapse Rate** (λ_i): The probability of random responses, modeled hierarchically as $\lambda_i \sim \text{Beta}(\alpha_\lambda, \beta_\lambda)$.

Thresholds and slopes were modeled hierarchically across participants. For the Single section, $\mu_i^{\text{Single}} \sim \mathcal{N}(\mu_{\text{group}}, \sigma_\mu)$ and $\log(\sigma_i^{\text{Single}}) \sim \mathcal{N}(\sigma_{\text{group}}, \sigma_\sigma)$. The Ensemble section included offsets for thresholds ($\Delta\mu$) and slopes ($\Delta\sigma$):

$$\mu_i^{\text{Ensemble}} = \mu_i^{\text{Single}} + \Delta\mu, \quad \sigma_i^{\text{Ensemble}} = \sigma_i^{\text{Single}} \cdot \Delta\sigma,$$

with $\Delta\mu \sim \mathcal{N}(\mu_{\Delta\mu}, \sigma_{\Delta\mu})$ and $\log(\Delta\sigma) \sim \mathcal{N}(\mu_{\Delta\sigma}, \sigma_{\Delta\sigma})$.

The primary parameter of interest was $\Delta\sigma$, representing the slope difference between sections, as it reflects whether

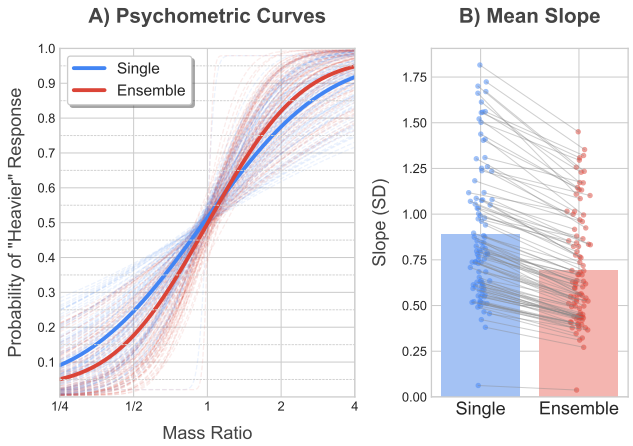

Figure 3: Experiment 1 results. (A) Solid lines represent average psychometric curves across participants for single and ensemble mass discriminations. Light dashed lines represent individual curves. (B) Changes in the estimated slope of the psychometric curve between the single and ensemble trials. Dots represent individual participants, with the bar representing the average.

participants demonstrated greater sensitivity in the Ensemble section.

The Bayesian model revealed that observers discriminated mass differences more accurately for Ensemble stimuli compared to the Single stimuli ($\Delta\sigma = \sigma_{\text{Ensemble}}/\sigma_{\text{Single}}$; mean: 0.76, 95% CrI: [0.69, 0.84]); see Fig. 3). We found a small difference between stimuli types in the threshold (mean: 0.039, 95% CrI: [0.00, 0.08]), only differing slightly from zero for Single stimuli (mean: -0.034, 95% CrI: [-0.066, -0.004]) but not for Ensemble stimuli (mean: 0.005, 95% CrI: [-0.020, 0.031]), suggesting judgments with little bias.*

Furthermore, the difference in psychometric slopes was reliable across individuals, with all participants being best explained by a steeper slope in the Ensemble versus the Single section, suggesting that the better discrimination for Ensemble stimuli is not just driven by extreme performance from a few participants.

Experiment 2

While Experiment 1 showed that participants can discriminate average mass more accurately when comparing ensembles of marbles than when comparing individuals, it is possible that participants were nonetheless aggregating information from individual marbles, but performing better in the Ensemble condition because more information was available. To rule out this alternative explanation—that ensemble judgments could arise entirely from aggregating individual-level detections—we designed a second experiment, explicitly testing whether participants' ensemble judgments exceed what could be explained by noticing changes in individual objects alone.

Following the paradigm of (Haberman & Whitney, 2011), we tested participants' ability to localize changes in an ensemble, as well as judgments of the ensemble as a whole. If people's ensemble judgments rely solely on item-level cues (i.e., identifying changed marbles), their overall mass discrimination should be predictable from item-level localization performance. However, if participants encode and compare the mean mass of the ensemble, they should outperform this item-level prediction.

Participants We collected data from 130 participants on Prolific, aiming for 80 participants after exclusions (see below). The experiment lasted approximately 20 minutes, and each participant was compensated $5 for their time.

Stimuli We used the same Ensemble stimuli as in Experiment 1 (sets of 25 marbles with average masses spanning from $\frac{1}{4}$ to 4 times the reference mass). Here, however, the test videos were not all compared to the same reference mass. Instead, the comparison videos showed modifications of the original test stimuli by taking five of the heaviest or lightest marbles and making their masses 5 times lighter or heavier respectively (changing the average mass of the entire ensemble by a factor of 1.4).

Procedure Participants completed 60 trials. On each trial, participants saw two videos and were first asked to judge whether the average marble mass in the second video was lighter or heavier than the first (as in Experiment 1), and then to select *one* of the marbles that had changed their weight between video 1 and video 2 by selecting its grid position on the cloth (see Fig 2c).

We included five catch trials in which 15 of the 25 marbles were altered by a factor of 16, yielding an obvious difference, changing the average mass by a factor of 10. Participants who did not correctly identify the heavier ensemble in at least 4 of these 5 trials were excluded from further analysis.

Analysis We tested whether participants' mass discrimination exceeded what would be expected if they relied purely on detecting which marbles changed (item-level detection). If item-level detection plus guessing fully explains mass discrimination, then participants' discrimination accuracy should be predicted by their localization performance. Conversely, if participants utilize an aggregate physical property (the ensemble's mean mass), their observed mass-discrimination accuracy should surpass what would be predicted by their item-level change localization accuracy. We analyze these results agnostic to how people might be extracting individual or ensemble information (e.g., by attending solely to cloth deformation) – instead our method tests solely whether judgments of ensemble masses can be explained by aggregating percepts of individual objects.

*A frequentist analysis – using a GLMM with a logistic link function and random slopes for threshold and slope for each individual – produced qualitatively similar results to the Bayesian analysis.

Results Overall, participants performed above chance in both tasks, but far from ceiling. The average participant selected the heavier ensemble 72.1% of the time (SD = 9.3%), and located an individually changed marble 42.7% of the time (SD = 12.1%).

Table 1: Total counts of correct vs. incorrect judgments for discrimination and localization across all trials and participants

	Discrimination Correct	**Discrimination Wrong**
Localization Correct	1674	204
Localization Wrong	1499	1023

We first analyzed our results following the method of Haberman and Whitney (2011). In our stimuli, an individual change is perfectly diagnostic of the change in ensemble mass, so we assume that if people localize the change they will get the discrimination correctly. Because participants correctly selected a changed marble in 42.7% of the trials, we can calculate the average probability of noticing any individual difference (p_{notice}) when accounting for guessing on the localization task as 28.4%, using the equation:

$$p_{\text{loc\_correct}} = p_{\text{notice}} + (1 - p_{\text{notice}}) \cdot 0.2 \tag{2}$$

If people use individual changes as the basis for their ensemble mass judgments, then they should correctly identify the heavier marble group, again accounting for chance, as:

$$p_{\text{mass\_correct}} = p_{\text{notice}} + (1 - p_{\text{notice}}) \cdot 0.5 \tag{3}$$

This would suggest that if participants were relying on information from individual marbles, they would have an accuracy of 64.2% on the ensemble discrimination task. However, their performance of 72.1% was statistically significantly greater than would be expected (3173/4400, exact binomial test, $p \approx 0$), suggesting that participants' ensemble judgments outperform what would be expected if they were simply aggregating individual information.

We can further look at the trials in which participants *failed* the change localization task, and so cannot be using individual information. We find that their accuracy is still reliably above chance (59.4%, 1499/2522, exact binomial test, $p \approx 0$). Thus, even when individual information is not available, participants still have enough information to correctly make judgments about the ensemble.

Bayesian Model: To further investigate the distinction between item-level and ensemble-level processes, we implemented a hierarchical Bayesian model (Fig. 4). This model estimates two key parameters for each participant:

1. **Notice Probability:** A latent probability representing how often participants detect the changed marbles on a given trial. When changes are detected, the model assumes perfect accuracy for both tasks (mass discrimination and localization). When changes are not detected, guessing is modeled with fixed probabilities of 0.5 for mass discrimination (2AFC) and 0.2 for localization (random chance among five marbles).

2. **Performance difference:** A participant-specific offset added only to the mass-discrimination probability. This parameter captures sensitivity exceeding what is explained by detecting changed marbles alone.

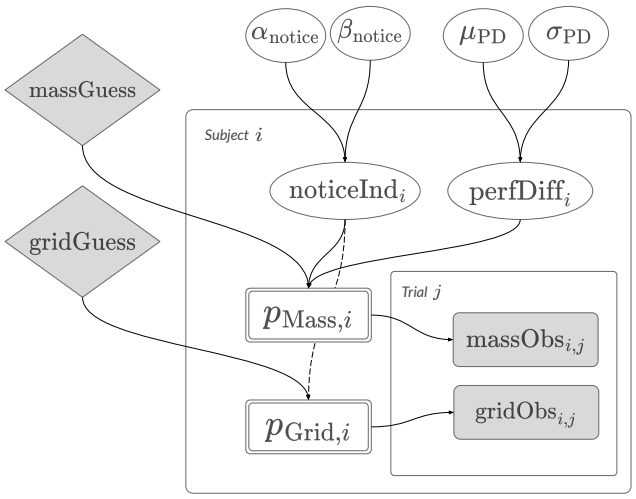

Figure 4: Graphical overview of the hierarchical Bayesian model for Experiment 2, which tests whether participants' mass discrimination judgments exceed what is predicted by detecting individual changes alone. The model estimates an additive performance difference (perfDiff) representing this extra ensemble-level advantage.

The posterior distribution of the performance difference parameter was analyzed to test whether it was credibly greater than zero. If the 95% credible interval for the performance difference parameter lies entirely above zero, this indicates a reliable ensemble-level advantage in mass perception.

The hierarchical Bayesian model estimated participants' probability of detecting changes in individual marbles (item-level detection) and their sensitivity to ensemble-level differences (see Fig. 4). We define the probabilities $p_{\text{Mass},i}$ and $p_{\text{Grid},i}$ for each subject i as:

$$p_{\text{Mass},i} = \text{noticeInd}_i + (1 - \text{noticeInd}_i) \cdot \text{massGuess} + \text{perfDiff}_i$$
$$p_{\text{Grid},i} = \text{noticeInd}_i + (1 - \text{noticeInd}_i) \cdot \text{gridGuess}$$

where noticeInd_i represents the probability of detecting individual changes, massGuess and gridGuess are baseline probabilities, and perfDiff_i captures the ensemble-level advantage.

The notice probability had a mean of 0.284 (SD = 0.009), indicating that the average participant would detect changes in individual marbles on approximately 28% of trials. Importantly, the performance difference parameter, which reflects sensitivity to ensemble-level properties beyond item-

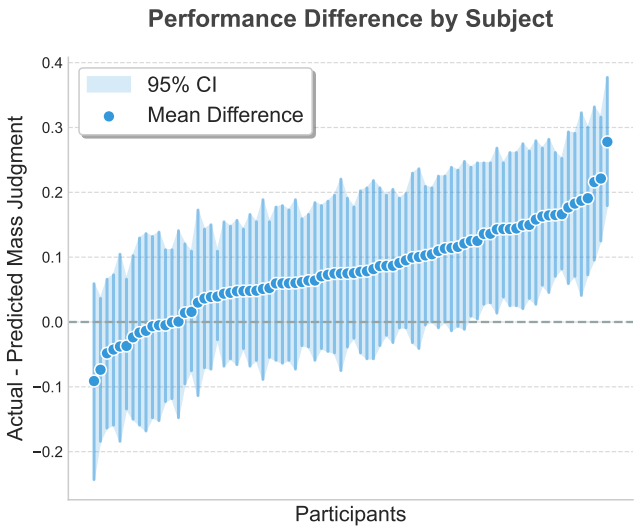

Figure 5: Difference between observed and predicted (from individual localization) ensemble mass discrimination accuracy by subject. Markers show the mean difference (observed minus predicted) with a 95% bootstrapped confidence interval. The zero line marks parity; values above zero indicate performance that exceeds predictions from individual item change detection.

level detection, had a mean of 0.078 and a 95% credible interval above zero ([0.061, 0.094]). The majority of individual participants (68 of 80) had positive performance differences (Fig. 5), suggesting the effect is not driven by just a few people. This provides strong evidence that participants could access ensemble-level mass representations independently of their ability to detect changes in individual items.

Discussion

Our results show that people can extract ensemble properties such as mass in a way that cannot be fully explained by item-by-item processing. In Experiment 1, participants judged ensemble masses more accurately than individual masses, suggesting that mass perception operates differently for groups of objects. In Experiment 2, participants' ability to compare the mean mass of two groups exceeded what would be predicted by detecting and tracking only the changed marbles. Even considering our stimuli, which were intentionally idealized to facilitate individual object processing, participants still seemed to treat the ensembles as a group rather than purely as a collection of individuals. This suggests that, similar to ensemble perception in other domains, the brain forms statistical representations of physical properties; but unlike size or orientation, ensemble mass perception requires observing objects interacting over time.

An open question is whether ensemble physics represents a distinct mode of processing or if it reflects a transition between reasoning about individual objects and reasoning about continuous substances like liquids. Some behavioral evidence suggests that humans process "stuff," like granular materials or non-solid substances, differently from discrete objects, often treating them as fluids (Kubricht et al., 2017; Bates, Yildirim, Tenenbaum, & Battaglia, 2019). Neuroimaging research further suggests that different brain regions are involved in perceiving granular and liquid substances versus solid objects (Paulun et al., 2023). Our findings suggest that ensemble physics might sit at an intermediate stage between these two modes of representation: the objects in our experiments remained discrete and separable but were processed collectively, as a group with emergent properties.

One factor influencing this transition might be the number and types of objects in a scene. It is possible that when only a few objects are present, people encode them individually, while at larger numbers, they switch to an ensemble-level representation. If so, there could be a threshold at which object-based reasoning gives way to ensemble processing. Additionally, spatial arrangement may play a role in whether a group is perceived as an ensemble. In our study, marbles were evenly distributed on a grid, reinforcing their treatment as individuals. Would the same effect hold if the objects were randomly distributed, tightly clustered, or mixed with objects of different sizes, colors, or textures? The extent to which ensemble perception generalizes across different configurations remains an open question.

While our study focused on mass, ensemble reasoning might extend to other physical properties. For example, people might extract summary representations of an ensemble's elasticity, friction, or density. Beyond aggregate properties that can be attached to individual objects, it is also worth considering whether people can extract relational representations from ensembles that do not exist for individuals. Viscosity, for instance, is defined by how a material resists deformation or flow, which depends on the movement of its individual components relative to one another. While people can perceive the viscosity of liquids (Bates et al., 2019; van Assen, Barla, & Fleming, 2018), can they analogously perceive relational properties of ensembles? This would suggest that they are computing group-level physical properties that emerge from the interaction of many elements. Can the perception of liquids even be thought of as a form of ensemble perception? Exploring these questions will further enhance our understanding of the representations and computations underlying intuitive physics and material perception.

Traditional models of cognition emphasize simulating individual objects. However, our results suggest that the mind may represent ensembles in a distinct manner. The brain may represent different classes of physical entities—individual objects, ensembles, and perhaps even "stuff", i.e., substances like liquids—each with its own computational properties. How ensemble representations are constructed, which properties they compute, and how they interact with individual-object representations remain open questions. Exploring these distinctions will improve our understanding of how the mind organizes and simulates the physical world.

Acknowledgments

This work was supported by NSF grant 2121009 awarded to JBT and KAS, and the German Research Foundation Project PA 3723/1-1 awarded to VCP.

References

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398. doi: 10.1111/j.1467-9280.2008.02098.x

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. doi: 10.1111/1467-9280.00327

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), e1007210.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753. doi: 10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734. doi: 10.1037/a0013899

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5), 855–859. doi: 10.3758/s13423-011-0125-6

Hespos, S. J., Ferry, A. L., Anderson, E. M., Hollenbeck, E. N., & Rips, L. J. (2016). Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological science*, *27*(2), 244–256.

Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S.-C., & Lu, H. (2017). Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Cogsci.*

Paulun, V. C., & Fleming, R. W. (2020). Visually inferring elasticity from the motion trajectory of bouncing cubes. *Journal of Vision*, *20*(6), 6–6.

Paulun, V. C., Pramod, R., & Kanwisher, N. (2023). Things versus stuff in the brain. *Journal of Vision*, *23*(9), 5096–5096.

Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of vision*, *17*(1), 20–20.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (2024). Intuitive physics as probabilistic inference. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Bayesian models of cognition : reverse engineering the mind.* Cambridge, MA: MIT Press.

van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current Biology*, *28*(3), 452-458.e4. doi: https://doi.org/10.1016/j.cub.2017.12.037

Watson, A. B. (2017). Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10. doi: 10.1167/17.3.10

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*, 105–129. doi: 10.1146/annurev-psych-010416-044232

Yildirim, I., Smith, K. A., Belledonne, M. E., Wu, J., & Tenenbaum, J. B. (2018). Neurocomputational modeling of human physical scene understanding.