Reconstructing mental representations of block towers: A new bias for physical stability in visual working memory

Stefan Uddenberg[1], Jun Kwak[2], & Brian J. Scholl[2]

[1] University of Illinois Urbana-Champaign
[2] Yale University

Running Head : Visual working memory and physical stability

Addresses for correspondence : Stefan Uddenberg
Department of Psychology
University of Illinois Urbana-Champaign
312 Psychology Building
Champaign, IL, 61820

Email : stefanu@illinois.edu

Word Count : 2943 (Intro/Results/Discussion) + 1111 (Methods)
Version : 8/14/26 — In press, *Psychonomic Bulletin & Review*

Abstract (210 words)

Recent studies have explored the perception of physical properties (such as mass and stability) in psychology, neuroscience, and AI, and perhaps the most popular stimulus from such studies is the *block tower* — since such displays (of stacked rectilinear objects) evoke immediate vivid impressions of physical (in)stability. Previous work has studied how people make higher-level decisions about such towers, often with explicit questions ("Which way will it fall?"), and limited to those tower properties that are systematically manipulated as explicit independent variables. Here, in contrast, we explored how such stimuli are spontaneously encoded into visual working memory in the first place. Using a novel drag-and-drop reconstruction task in a custom 3D interface, observers briefly viewed randomly generated block towers, and then simply reproduced them from memory — without ever being asked about stability, per se. By simulating both the original and reconstructed towers in a physics engine, we discovered a novel bias: observers reliably remember towers as more stable than they originally were, even when the reconstructions are otherwise accurate. This reveals for the first time how visual memory is structured (and biased) not only by how scenes are segmented into objects, but also by how those objects are physically related to each other — all without ever asking anyone anything about physics.

Keywords

Intuitive Physics; Visual Working Memory; Physical Stability; Block towers

1. Introduction

Take a moment to look at the images in Figure 1. What do you see? Of course you can appreciate the various low-level properties of those images, such as the colors and orientations of the various objects. But you may also readily perceive an even more obvious feature: each image depicts a stack of objects, evoking a vivid impression of physical (in)stability.

Figure 1. Familiar examples of physical (in)stability.

A vibrant and growing body of research in recent years has explored how we make judgments of physical stability when viewing such stimuli — including work in psychology (e.g. Battaglia et al., 2013; Hamrick et al., 2018; Mitko & Fischer, 2020), neuroscience (e.g. Fischer et al., 2016; Gallivan et al., 2014; Pramod et al., 2022; Schwettmann et al., 2019), and AI (e.g. Buschoff et al., 2025; Groth et al. 2018; Lerer et al. 2016; Sanchez-Gonzalez et al., 2018; Toussaint et al., 2018). And nearly all of this work explores human judgments of (in)stability in the most direct (and most explicit) way possible — by simply asking people overt questions about such displays, such as “Will this tower fall?”, “What direction will it fall in?”, or “How stable does this tower look, on a scale from 1 to 10?” (e.g. Battaglia et al., 2013).

1.1. From Judgment to Memory

Existing work on intuitive physics has explored several possible underlying mechanisms for such stability judgments, such as the notion that we have a ‘mental game engine’ that computes physical relationships to predict how dynamic scenes will unfold (Battaglia et al., 2013; Pramod et al., 2022; Ullman et al., 2017). But such proposals have mostly been agnostic or silent as to the roles of specific underlying cognitive processes in fueling such judgments, such as attention or spatial cognition (see also Firestone & Scholl, 2016; Mitko & Fischer, 2020).

Accordingly, the current work explores (for the first time, to our knowledge) how intuitive physics may structure — and bias — representations held in visual working memory (VWM). Of course, VWM doesn’t operate over raw images, per se, but rather stores structured object-based representations. For example, it has been claimed that VWM capacity limits are sometimes determined not by the brute amount of to-be-remembered information, but by how that information is distributed among a number of discrete segmented objects (e.g. Adam et al., 2017; Luck & Vogel, 1997; Ngiam et al., 2023; Vogel et al., 2001; but see Oberauer & Eichenberger, 2013; for reviews see Fukuda et al., 2010; Luck & Vogel, 2013; Mance & Vogel, 2013; Schurgin, 2018; Suchow et al., 2014; van den Berg et al., 2014). In this context, the present work seeks to push this sophistication even further: might VWM be structured (and biased) not only by how scenes are segmented into discrete objects, but also by how those objects are physically related to each other?

1.2. The Current Study: A Novel Reconstruction Task

We explored interactions between physical stability and VWM using a novel reconstruction task. Observers first briefly viewed a randomly generated block tower (consisting of only a few colored blocks, as in the examples from Figure 2).

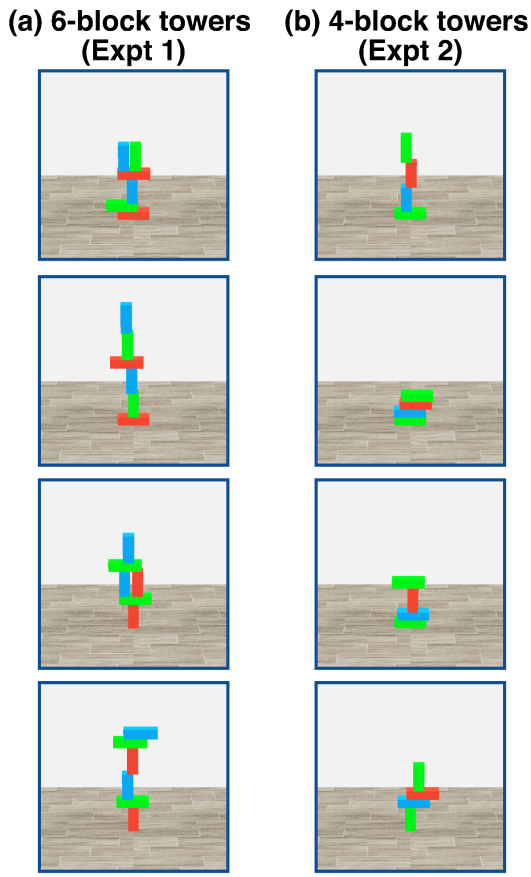

Figure 2. Examples of simulated block towers from our experiments: (a) 6-block towers, from Experiment 1, (b) 4-block towers, from Experiment 2.

After that tower had disappeared, observers then had to immediately reconstruct it from memory, using a novel virtual interface, with its constituent blocks initially spread out along the “floor” before them (as in Figure 3).

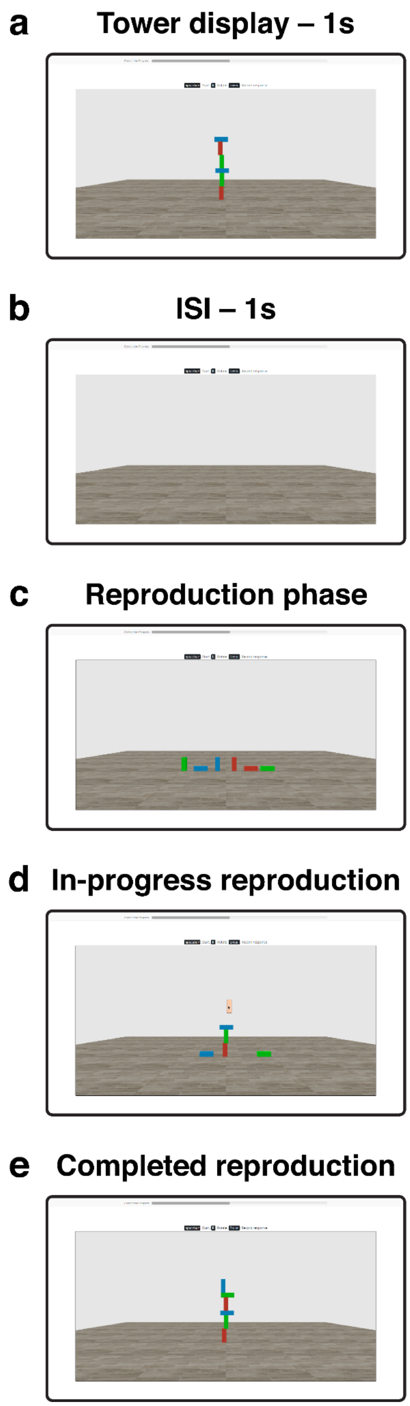

Figure 3. Depiction of the tower reconstruction task for a sample six-block tower. (a) The initial presentation. (b) A brief pause. (c) The reconstruction task begins with the component blocks all arrayed along the ground plane. (d) Here the subject is roughly halfway through reconstructing the tower, and is about to place the red block onto the blue block. (e) The subject's final reconstruction.

This deceptively simple experimental method has two key advantages in the current context. First, whereas studies of explicit stability judgments must decide beforehand which relevant features to vary (e.g. the towers' sizes, symmetries, shape envelopes, etc.), the current method has no such constraints: the towers are initially randomly generated, and in principle VWM for such stimuli could be biased in many different (multidimensional) ways. As such, simple reconstruction allows for data-driven insight, without needing to pre-specify the nature or extent of any memory biases. Second, this experimental paradigm can readily proceed without ever overtly mentioning physical stability — whereas such properties have been inextricably central to the explicit questions used in prior work. And as such, this task may have more generalizability, as it does not draw observers' attention to features (e.g. stability) that they might not otherwise prioritize.

2. Experiment 1: VWM for 6-Block Towers

In this first test of visual working memory for physical stability, we employed the novel reconstruction task with images of towers randomly constructed from six blocks (as in Figure 2a).

2.1. Materials & Methods

3.1.1. Participants. We recruited 100 U.S.-based subjects (32 females; mean age=30.27, *SD*=10.37) using the Prolific Academic online labor market. (For discussion of this pool's nature and reliability, see Peer et al., 2017, 2021). This sample size was preregistered, and was chosen arbitrarily to be in line with other recent studies of intuitive physics (e.g. Bi et al., 2025; Hamrick et al., 2018). 12 additional subjects were excluded for one or more of the following preregistered reasons: reporting that they were inattentive during the experiment, as described in more detail below (3); clicking away from the experiment window on one or more reconstruction trials (5); stating that they had already participated in the study (3); or having produced one or more towers that were less than 4 units tall (2).

2.1.2. Apparatus. The experiment was conducted using custom software written using a combination of Python, Javascript, CSS, and HTML. Key libraries used included jsPsych (de Leeuw, 2015) for experiment presentation, and babylon.js (https://www.babylonjs.com) for 3D scene creation. Subjects completed the experiment via a custom web page which could be loaded in any modern web browser on their own laptop or desktop computers. (Because our experiment required a relatively large display and the use of a computer mouse to make responses, mobile devices such as phones and tablet computers were explicitly disallowed, and attempts to access the experiment from such a device led to its immediate termination along with an error message.) The experiment window expanded to take up the full extent of the subject's screen after they consented to participate. We also included a link to a simple custom browser compatibility check, allowing each subject the opportunity to ensure that their system was able to process the sorts of 3D scenes that would be shown.

2.1.3. Stimuli. The kinds of stimuli subjects saw and reproduced are depicted in Figure 2a. Each randomly generated tower shown to subjects consisted of six colored blocks, with each block measuring 1×1×3 units in size. (Because subjects completed the experiment on their own browsers and displays, we report all stimulus dimensions in arbitrary units. On a typical 1600px×900px 15.6-inch laptop display viewed from 57cm, each block subtended approximately 1.53° by 0.56° when horizontal.) The towers were constructed with blocks that could either be oriented horizontally or vertically, and were built up such that they were "locally stable" (so that whenever a new block was added to the tower, it would not fall off of its immediate support in isolation, based on the position of its center of gravity). Each of the 3D matte blocks were colored either red ([0, 74, 54]), green ([0, 255, 23]), or blue ([10, 176, 255]), and cycled through those colors in that order as the tower was built up from the ground. The towers were shown in a simple virtual environment with a wooden textured floor and white walls (as in Figure 3). The scene was lit with an ambient white light source aiming up and to the left with an intensity of 0.25, and two white point light sources placed around the space (XYZ coordinates: [-20, 20, -20] and [20, -20, 20] respectively). The camera was positioned approximately 13.6 lengthwise-blocks away from and above the block tower scene (XYZ coordinates: [0, 8, -40]), and was angled to face slightly above the scene's origin (focused on the XYZ coordinate [0, 8, 0]).

2.1.4. Procedure. Subjects first gave their informed consent and agreed to answer open-ended questions, using a procedure designed to reduce subject attrition (Zhou & Fishbach, 2016). The experiment window then expanded to take up the full extent of the screen, and subjects read detailed written instructions for the task, as described below. On each of two trials, subjects were shown the empty virtual environment. After a keypress, a randomly generated block tower was displayed for 1 second (Figure 3a). And after another second of again viewing the empty virtual environment (Figure 3b), the tower's constituent blocks reappeared on the virtual floor, equally spaced apart in a random order and in random orientations (horizontal or vertical; Figure 3c). Subjects were instructed beforehand to reconstruct the tower that they just saw by clicking and dragging the blocks into position. They were able to change a block's orientation between horizontal and vertical by pressing the "R" key on the keyboard while dragging it. When the subject released a block, it instantly dropped straight down until it collided with either the ground or with another block, at which point it stopped, remaining fixed in place regardless of its stability. Subjects adjusted the blocks as often (and for as long) as they wished, after which they pressed the spacebar to submit their reconstruction (which they were only allowed to do after moving at least 5 of the blocks from their initial starting positions). A sample in-progress reconstruction is depicted in Figure 3d, and a sample final reconstruction is depicted in Figure 3e.

Subjects were then asked a series of debriefing questions, including: (a) what they thought we were testing; (b) how well they thought they did in the reconstruction task, on a scale of 1 (terrible) to 100 (perfect); (c) whether they used any particular strategies in the task;

(d) whether they interrupted their completion of the survey in any way (e.g. by refreshing the task page or switching tabs/windows); (e) whether any part of the procedure was unclear or if they encountered any problems during the experiment; (f) whether they had participated in the task before; and (g) how well they were able to pay attention over the course of the experiment, on a scale of 1 (very distracted) to 100 (very focused). After submitting their responses, they were redirected to Prolific for compensation, and the experiment ended.

To confirm that subjects reproduced the towers with sufficient fidelity, we performed a supplemental memory analysis comparing silhouette overlap between the initially presented and reproduced towers, compared to randomly permuted pairs of towers (included in tabs 14 and 15 of the Supplementary Data File, with details for both experiments).

2.2. Results

We measured tower stability offline by simulating the influence of gravity for 10 seconds (600 frames) on both the towers initially shown to subjects (initial towers), and the towers that they reconstructed (reconstructed towers). Simulations were generated using the Panda3D game engine (https://www.panda3d.org/) in concert with the Bullet physics engine (https://pybullet.org/wordpress/), with each block assigned a mass of 1 unit and with gravity set to –9.81.

Per our preregistered analysis plan, we first evaluated the stability of the initial towers and reconstructed towers categorically. For this purpose, we simply categorized a tower as stable if (during its subsequent simulation) no block moved more than 0.5 units of distance from its original starting location, or as unstable if at least one block moved more than this distance. And we further quantified instability in terms of the brute number of blocks that moved more than this distance (from 1 to 6). For stable towers (N=144), we were thus able to measure how many reconstructions became less stable, and how many stayed stable. (Note that an initially stable tower can never become more stable via this initial categorical metric.) And for unstable towers (N=56), we were able to measure how many reconstructions became less stable, how many became more stable, and how many remained at the same level of instability. The resulting percentages are depicted for initially stable towers in Figure 4a, which reveals that most stable towers were reproduced as stable (by approximately a 2 to 1 ratio — 98 stable vs. 46 less stable — with this difference being highly reliable by a binomial test, p<.001). More theoretically consequential were the resulting percentages for the initially unstable towers, as depicted in Figure 4b. Here we observed a striking and unexpected pattern, with reconstructed towers being reliably more stable than were the initial towers (as confirmed by a difference of observed proportions, as measured in terms of the raw counts — 36 more stable vs. 11 unchanged vs. 9 less stable — $X^2[2, N=56]=24.25$, $p<.001$, $\phi=0.66$). This effect was also dramatic in its magnitude, with reconstructions becoming more stable 4 times as often as they became less stable.

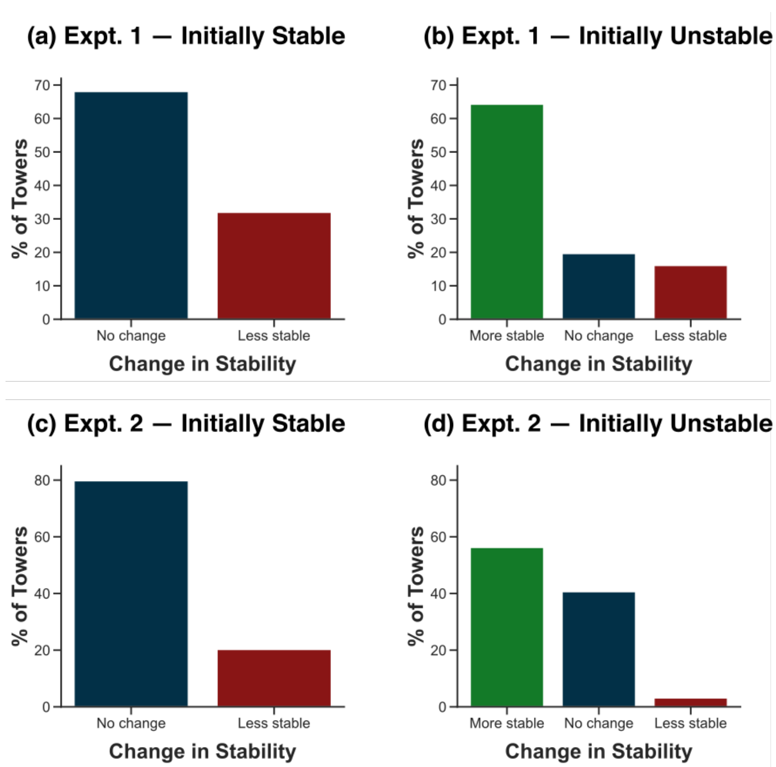

Figure 4. Categorical changes in stability in reconstructed towers (compared to initial towers) in Experiment 1 (a and b) and Experiment 2 (c and d).

Beyond these categorical measures, we also analyzed the reconstructed towers in terms of three continuous measures of stability, each of which compared the initial scenes to the final post-simulation tableau. The first of these was the same metric used in the initial categorical analyses: the number of blocks that moved more than 0.5 units of distance from their original starting locations. In addition, we also measured both the dispersion of the blocks (calculated as the difference between the x-coordinates of the blocks that were furthest from one another in the final tableau) and the average distance that blocks moved (from the initial scene to the final tableau), when they moved at all (defined as moving more than 0.001 units of distance from their starting location). These three continuous metrics of (in)stability are depicted in Figure 5 in terms of histograms for the changes between initial unstable towers and their reconstructions. In each case, values that indicated greater stability in the reconstructed towers are colored green, values that indicated less stability are colored red, and a single bar closest to the same level of stability is colored blue. Inspection of this graph reveals a clear and consistent pattern: for all three metrics, subjects reliably reconstructed towers as more stable than they initially were. This impression was confirmed with a series of t-tests: relative to the initial towers, reconstructions had fewer blocks move (Figure 5a; $M=-2.23$, $SD=2.66$; $t(55)=6.27$, $p<.001$, $d=0.84$), had less dispersion (Figure 5b; $M=-8.07$, $SD=8.97$; $t(55)=6.74$, $p<.001$, $d=0.90$), and had less average movement among those blocks that did move (Figure 5c; $M=-4.07$, $SD=3.96$; $t(55)=7.70$, $p<.001$, $d=1.03$).

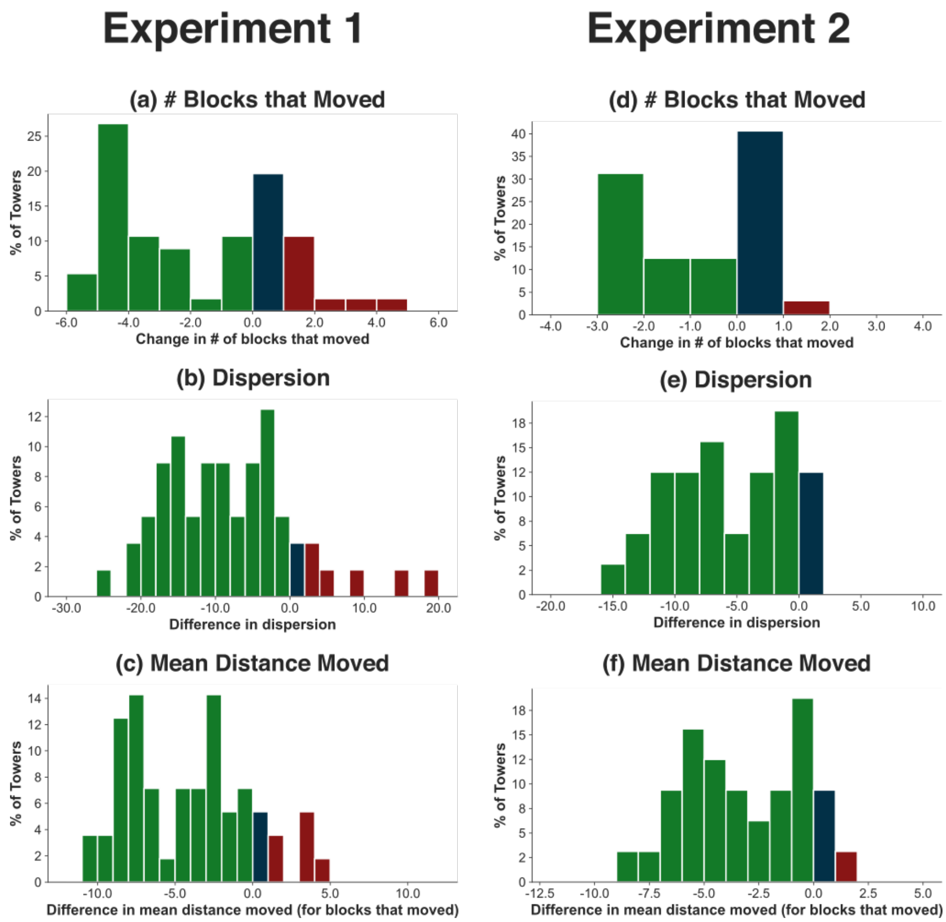

Figure 5. Histograms for changes between initial unstable towers and their reconstructions, for both Experiments 1 and 2, in terms of (a and d) number of blocks that moved, (b and e) the resulting dispersion, and (c and f) the average distance moved (for blocks that did move). Values that indicated greater stability in the reconstructed towers are colored green, values that indicated less stability are colored red, and a single bar closest to the same level of stability is colored blue.

3. Experiment 2: VWM for 4-Block Towers

Although each trial in Experiment 1 involved only a single tower, that tower was always composed of six different blocks, with this value being above most contemporary estimates on the limits of short-term memory (as in the ‘magic number 4’; e.g. Cowan, 2001). As such, the bias to reproduce towers as more stable than they originally were might have stemmed not from a memory bias per se, but rather from an inference about likely degrees of stability in the absence of a robust memory. To test this possibility, we replicated Experiment 1 but now using towers composed of only four blocks each — where this less demanding load was more comfortably within the range of typical memory capacity estimates.

3.1. Materials & Methods

This experiment was identical to Experiment 1, except as noted. We again recruited 100 U.S.-based subjects (48 females; mean age=32.06, *SD*=11.48) using Prolific, with this preregistered sample size chosen to match that of Experiment 1. 10 additional subjects were excluded for one or more of the following reasons: reporting that they were inattentive during the experiment (5); clicking away from the experiment window on one or more reconstruction trials (4); or having produced one or more towers that were less than 3 units tall (2). Each of the blocks now cycled through the colors green, blue, and red (in that order) as the tower was built up from the ground.

3.2. Results

The percentages of towers that changed their categorical levels of stability are shown in Figure 4, broken down by towers that were initially stable after simulation (with no blocks that moved; Figure 4c) vs. those that were initially unstable (with 1 or more blocks that moved; Figure 4d). Inspection of Figure 4c again shows that most stable towers (N=168) were reproduced as stable (by approximately a 4 to 1 ratio — 134 stable vs. 34 less stable — with this difference being highly reliable by a binomial test, $p<.001$). Critically, we also replicated our original striking pattern for initially unstable towers (N=32), as depicted in Figure 4d: reconstructed towers were reliably more stable than were the initial towers (as confirmed by a difference of observed proportions, as measured in terms of the raw counts — 18 more stable vs. 13 unchanged vs. 1 less stable — $X^2(2, N=32)=14.31$, $p<.001$, $\phi=0.67$). And now this effect was even more dramatic in its magnitude, with reconstructions becoming more stable 18 times as often as they became less stable.

Our three continuous metrics of (in)stability are depicted in Figure 5, again in terms of histograms for the changes between initial unstable towers and their reconstructions. Inspection of this graph reveals the same clear and consistent pattern observed in Experiment 1: for all three metrics, subjects reliably reconstructed towers as more stable than they initially were. This impression was again confirmed with a series of t-tests: relative to the initial towers, reconstructions had fewer blocks move (Figure 5d; M=–1.28, SD=1.37; $t(31)$=5.28, p<.001, d=0.93), had less dispersion (Figure 5e; M=–5.62, SD=4.62; $t(31)$=6.88, p<.001, d=1.22), and had less average movement among those blocks that did move (Figure 5f; M=–3.17, SD=2.64; $t(31)$=6.80, p<.001, d=1.20).

4. Discussion

The central result of this study was a novel bias in visual working memory related to intuitive physics: observers reliably remembered block towers as more stable than they originally were, as measured in a reconstruction task. Beyond the statistical support for this conclusion as summarized in Figures 4 and 5, we can also see several representative examples of this in practice in Figure 6: in each case, the left column depicts an initial random tower shown to a subject, and the right column depicts their (more stable) reconstruction. This effect seemed especially robust in at least 6 related ways. First, the effect was highly statistically reliable, to a degree far below conventional significance thresholds — with the average p-value of the primary comparison across the experiments being p=.00039. Second, the magnitude of the primary comparison was also large, with observations in one of the experiments exhibiting this pattern by an 18:1 ratio. Third, we observed this same effect across many different measures of stability, including quantitative measures (such as the ultimate dispersion of the blocks after simulation). Fourth, this effect generalized to towers that both did and did not transcend typical estimates of VWM capacity. Fifth, this bias occurred despite subjects' reproductions being overall highly accurate (as detailed in tabs 14 and 15 of the Supplemental Data File for both experiments). Finally, this effect occurred despite subjects never being asked, told, or shown anything at all about the physical properties of the towers (in contrast to existing work in the published behavioral literature).

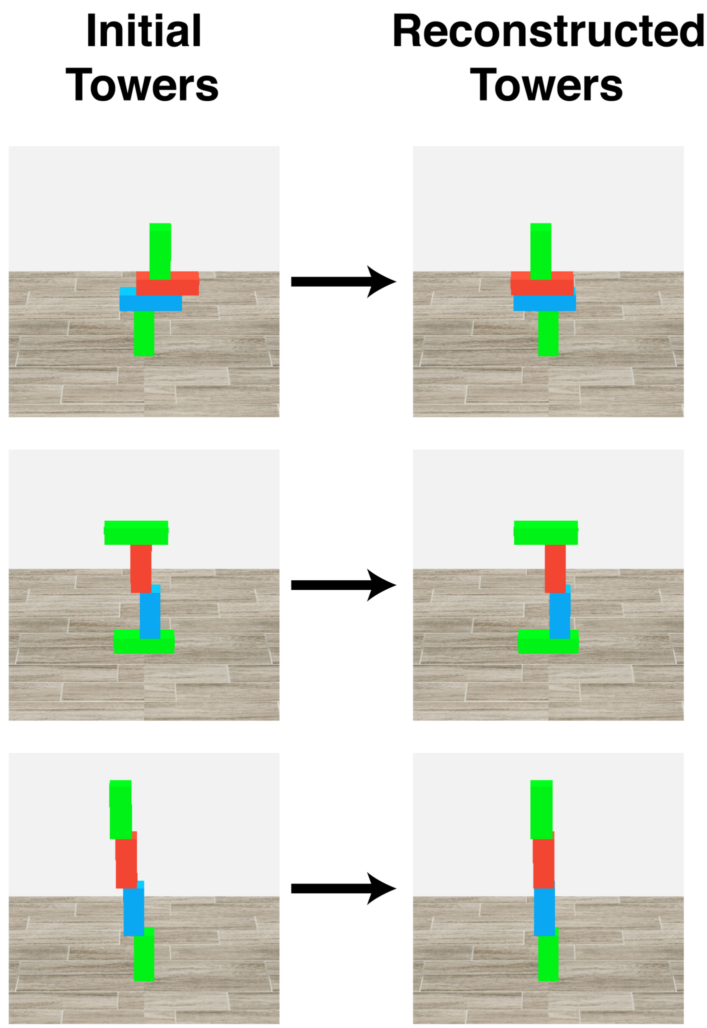

Figure 6. Examples of towers initially shown to subjects (left column) and subsequently reproduced as more stable (right column).

4.1. A Surprising Outcome?

We were surprised by this memory bias. In fact, we initially predicted the opposite outcome — that unstable towers would be misremembered as *more* unstable, exacerbating a key salient physical property of the towers — for at least 3 reasons:

First, past work suggests that instability may be especially attention-grabbing: it can not only guide selective attention (Firestone & Scholl, 2016) and lead to higher levels of perceived 'interestingness' of towers (Holdaway et al., 2021), but can even fuel a powerful visual search asymmetry: people are quicker to find less stable images among stable distractors (e.g., a tilted vase among upright vases) than stable images among less stable distractors (e.g., an upright vase among tilted vases; Yang & Wolfe, 2020).

Second, note that the memory bias we actually observed leads to the mitigation (or removal) of a distinctive visual feature (viz. instability), whereas past work has more often observed augmentation (or addition) of salient features. Perhaps the most famous and emblematic examples of this are caricature effects for faces, wherein exaggerated representations of a known person are more easily and quickly recognized than are veridical depictions (e.g. Lee et al., 2000; Rhodes, 1997). Such caricature effects may be a domain-general feature of perception (Sun et al., 2024), as they are found in many contexts including in the auditory domain (Whiting et al., 2020).

And third, salient features are also over-weighted in many other contexts, including the extraction of temporal summary statistics — as when looming circles bias estimates of mean size (Albrecht & Scholl, 2010), or more intense emotional expressions bias estimates of mean emotion (Goldenberg et al., 2022). In the face of such evidence, the robust demonstration of a bias for *greater* physical stability in memory seems especially striking.

4.2. Advantages of Reconstruction

This work also demonstrates the utility of our (deceptively simple) reconstruction task: by having observers simply view and reconstruct towers from memory, we were able to uncover a novel memory bias that had not been observed in previous work. This approach allowed us to avoid two major downsides of most prior studies of intuitive physics.

First, whereas past work most often asked subjects directly about physical properties ("Will it fall?"), this reconstruction task allowed us to (implicitly) ask about mental representations of physical stability, without ever (explicitly) asking about physical stability. Second, by having subjects freely reconstruct the towers, there was no need for us as experimenters to choose which features of the towers we would manipulate or measure in advance: both the initially presented and reconstructed towers could vary with relatively few constraints on how they were built.

Future work could use this same approach to explore the underlying perceptual roots of various other intriguing phenomena, such as the perception of relationships between objects (e.g. containment or attachment; Hafri & Firestone, 2021; Strickland & Scholl, 2015) or causal history (Chen & Scholl, 2016; Schmidt et al., 2016).

4.3. Why?

Where does this memory bias come from? One deflationary possibility that our existing data can refute is simple familiarity. In principle, our findings could reflect a bias *for* stability rather than *towards* stability. That is, subjects may be predisposed to produce towers that are entirely, categorically stable — perhaps due to the ubiquity of such arrangements in everyday life, wherein unstable arrangements don't tend to stick around for long. (If you see a group of objects such as that in the upper left quadrant of Figure 1, it will probably look different the next time you see it.) This account, however, predicts that initially unstable towers should typically be reconstructed as entirely stable. But this was not the case: subjects tended to reconstruct unstable towers as *more stable*, even when the result was still categorically *unstable*. This can be appreciated both from the roughly continuous distributions seen in the histograms of Figure 5, and in Figures S1 (depicting the changes in stability for each of the towers, across our continuous metrics of stability, for both experiments) and S2 (depicting the tremendous overall reproduction accuracy across both experiments) in tabs 13 and 16 of the Supplementary Data File. Our results thus reflect a bias towards (but not categorically for) stability. Indeed, even if we limit our analyses solely to the most unstable towers, we can see that participants were quite willing to reproduce even the most initially unstable towers as very unstable — just a bit less so than in the original displays — thus ruling out the possibility that they were merely hesitant to produce unstable configurations (see Figures S3 and S4 in tabs 17 and 18 of the Supplementary Data File; these are analogous to Figures 4 and 5 shown earlier in the manuscript, but focused only on the most initially unstable towers).

Another deflationary possibility is that participants were not "reconstructing" the towers, but merely "constructing" them anew — creating brand new towers cut from whole cloth, as if they had closed their eyes during the main task. However, our key analyses all hold up even when constrained to just those towers that were reproduced well, as operationalized by good silhouette overlap (from a side-on view) between the initially presented and reconstructed towers (see tab #19 of the Supplementary Data File for full details of these analyses for both experiments).

A final alternative account might appeal to the possibility that participants reduced the complexity and details of the towers by aligning the blocks more, which would result in greater stability only as a "side effect" of a more general response tendency to produce simpler towers. However, our results hold even when we limited our analyses to those towers that showed no more alignment of the blocks in the reproductions (see Supplementary Data File tab #20).

Having dispatched these deflationary possibilities, we can also identify at least three theoretically interesting reasons for the bias towards physical stability in visual working memory. One possibility is that it stems from an urge to *improve* or shore up the towers. If you saw the arrangement from the upper left quadrant of Figure 1 on your desk, for example, your next move might be to manually align the coins by hand. That such an effect would emerge even when unnecessary (and even with virtual objects on a computer screen!) may speak to how ingrained such impulses can be. Another possibility is that the effect derives from memory biases of the local stability of each block, as in past findings of hierarchical encoding of visual ensembles (Brady & Alvarez, 2011). In effect, the global tendency towards more stability may be driven by ensemble coding of local stability — since the initial towers all had entirely locally stable configurations. Finally, it is intriguing to consider the possibility that people effectively have a prior expectation for stability — not in the form of a conscious belief, but rather as a type of spontaneous Bayesian prior in perception and memory. We know from past work, for example, that such priors can readily bias downstream reproductions in other contexts (Langlois et al., 2021; Uddenberg & Scholl, 2018). And these priors may have deep connections to the classic Gestalt notion of “Prägnanz” (Van Geert & Wagemans, 2023; Van Geert & Wagemans, 2024). Whichever of these (non-mutually exclusive) possibilities may lie behind the current results, this project demonstrates a new way in which two of the most active research areas in our field — visual working memory, and intuitive physics — may have unexpectedly rich connections.

Open Practices Statement

The hypotheses, methods, and analysis plan were preregistered prior to data collection. Preregistration for Experiment 1 can be found at https://aspredicted.org/FKX_7J7. Preregistration for Experiment 2 can be found at https://aspredicted.org/X7V_HJL. There were no deviations from the preregistration. All study materials, primary data, and analysis scripts are available in the Supplemental Data File.

Declarations

Acknowledgments: For helpful conversation, we thank Pranava Dhar, Alon Hafri, Aalap Shah, Kim Wong, Ilker Yildirim, and the members of the Yale Perception & Cognition Laboratory.

Funding: This project was funded by ONR MURI #N00014-16-1-2007 awarded to BJS.

Conflicts of interest/Competing interests: The authors declare no conflicts of interest.

Ethics approval: This study was performed in line with the principles of the Declaration of Helsinki. All experimental methods and procedures were approved by the Yale University Institutional Review Board.

Consent to participate: Informed consent was obtained from all individual subjects included in the study.

Consent for publication: Not applicable

Availability of data and materials: The raw data for each experiment can be viewed in the online supplementary data file for this study. Preregistrations for each experiment can be viewed at: https://aspredicted.org/FKX_7J7 (Experiment 1), and https://aspredicted.org/X7V_HJL (Experiment 2).

Code availability: Analysis scripts can be viewed in the online supplementary data file for this study.

Authors' contributions: **Stefan Uddenberg**: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Jun Kwak**: Conceptualization, Methodology, Software, Writing - Review & Editing. **Brian J. Scholl**: Conceptualization, Project Administration, Funding Acquisition, Supervision, Writing - Original Draft, Writing - Review & Editing

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97.

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, *21*, 560–567.

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*, *110*, 18327–18332.

Bi, W., Shah, A. D., Wong, K. W., Scholl, B. J., & Yildirim, I. (2025). Computational models reveal that intuitive physics underlies visual processing of soft objects. *Nature Communications, 16*:6303, 1–15.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*, 384–392.

Buschoff, L., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in multimodal large language models. *Nature Machine Intelligence, 7*, 96–106.

Chen, Y.-C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*, 923–930.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12.

Firestone, C., & Scholl, B. (2016). Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision*, *16*(12), Article 689.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*, E5072–E5081.

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology*, *20*, 177–182.

Gallivan, J. P., Cant, J. S., Goodale, M. A., & Flanagan, J. R. (2014). Representation of object weight in human ventral visual cortex. *Current Biology*, *24*, 1866–1873.

Goldenberg, A., Schöne, J., Huang, Z., Sweeny, T. D., Ong, D. C., Brady, T. F., Robinson, M. M., Levari, D., Zaki, J., & Gross, J. J. (2022). Amplification in the evaluation of multiple emotional expressions over time. *Nature Human Behaviour*, *6*, 1408–1406.

Groth, O., Fuchs, F. B., Posner, I., & Vedaldi, A. (2018). ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 702–717).

Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, *25*, 475–492.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1773–1778).

Holdaway, C., Bear, D. M., Radwan, S. F., Frank, M. C., Yamins, D. L. K., & Fan, J. E. (2021). Measuring and predicting variation in the interestingness of physical structures. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Langlois, T. A., Jacoby, N., Suchow, J. W., & Griffiths, T. L. (2021). Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, *118*, e2012938118.

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, *11*, 379–385.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. In *Proceedings of The 33rd International Conference on Machine Learning* (pp. 430–438).

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*, 391–400.

Mance, I., & Vogel, E. K. (2013). Visual working memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 179–190.

Mitko, A., & Fischer, J. (2020). When it all falls down: The relationship between intuitive physics and spatial cognition. *Cognitive Research: Principles and Implications*, *5*, 1–13.

Ngiam W., Loetscher, K., & Awh, E. (2023). Object-based encoding constrains storage in visual working memory. *Journal of Experimental Psychology: General, 153*, 86–101.

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition*, *41*, 1212–1227.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Peer, E., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*, 1643–1662.

Pramod, R. T., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *eLife*, *11*:e71736.

Rhodes, G. (1997). *Superportraits: Caricatures and Recognition*. Psychology Press.

Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., & Battaglia, P. (2018). Graph networks as learnable physics engines for inference and control. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 4470–4479).

Schmidt, F., Spröte, P., & Fleming, R. W. (2016). Perception of shape and space across rigid transformations. *Vision Research*, *126*, 318–329.

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics*, *80*, 1035–1056.

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*:e46619.

Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*, 570–580.

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, & Psychophysics*, *76*, 2071–2079.

Sun, Z., Han, S., & Firestone, C. (2024). Caricaturing shapes in visual memory. *Psychological Science*, *35*, 722–735.

Toussaint, M. A., Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2018). Differentiable physics and stable modes for tool-use and manipulation planning. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation.

Uddenberg, S., & Scholl, B. J. (2018). Teleface: Serial reproduction of faces reveals a whiteward bias in race memory. *Journal of Experimental Psychology: General*, *147*, 1466–1487.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*, 649–665.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*, 124–149.

Van Geert, E., & Wagemans, J. (2023). What good is goodness? The effects of reference points on discrimination and categorization of shapes. *Journal of Experimental Psychology: Human Perception and Performance*, *49*(8), 1180–1201.

Van Geert, E., & Wagemans, J. (2024). Prägnanz in visual perception. *Psychonomic Bulletin & Review*, *31*(2), 541–567.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.

Whiting, C. M., Kotz, S. A., Gross, J., Giordano, B. L., & Belin, P. (2020). The perception of caricatured emotion in voice. *Cognition*, *200*, 104249.

Yang, Y.-H., & Wolfe, J. M. (2020). Is apparent instability a guiding feature in visual search? *Visual Cognition*, *28*, 218–238.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*, 493–504.