The 'Blindfold Test': Helping to Decide whether an Effect Reflects Visual Processing or Higher-Level Judgment

Benjamin van Buren[1] & Brian J. Scholl[2,3]

[1]Department of Psychology, The New School
[2]Department of Psychology, Yale University
[3]Wu-Tsai Institute, Yale University

Running Head	:	The 'Blindfold Test'
Addresses for correspondence	:	Benjamin van Buren Department of Psychology The New School New York City, NY, 10011
Email	:	vanburenb@newschool.edu
Phone	:	(914) 282 5593
Word Count	:	6474
Version	:	11/7/24 — In press, *Attention, Perception, & Psychophysics*

Abstract (241 words)

Experimenters often ask subjects to rate displays in terms of high-level visual properties, such as animacy. When do such studies measure subjects' visual impressions, and when do they merely reflect their judgments that certain features *should* reflect animacy? Here we introduce the 'Blindfold Test' for helping to evaluate the evidence for whether an effect reflects perception or judgment. If the same effect can be obtained not only with visual displays, but also by simply *describing* those displays, then subjects' responses may reflect higher-level reasoning rather than visual processing — and so other evidence is needed in order to support a 'perceptual' interpretation. We applied the Blindfold Test to three past studies in which observers made subjective reports about what they were seeing. In the first two examples, subjects rated stimuli in terms of high-level properties: animacy, and physical forces. In both cases, the key findings replicated even when the visual stimuli were replaced with (mere) descriptions, and we conclude that these studies cannot by themselves license conclusions about perception. In contrast, a third example (involving Motion-Induced Blindness) passed the test: subjects produced very different responses when given descriptions of the displays, compared to the visual stimuli themselves — providing compelling evidence that the original responses did not merely reflect such higher-level reasoning. The Blindfold Test may thus help to constrain interpretations of the mental processes underlying certain experimental results — especially for studies of properties that can be apprehended by both seeing and thinking.

Keywords: Perception, Cognition, Causal perception, Perception of animacy, Motion-induced blindness

Introduction

Psychologists have long been inspired by the possibility that visual processing traffics not only in low-level features such as motion and orientation, but also in seemingly higher-level properties such as causality (e.g. Michotte, 1946/1963; for a review see Scholl & Tremoulet, 2000), animacy (e.g. Heider & Simmel, 1944; for a review see Scholl & Gao, 2013), and causal history (e.g. Chen & Scholl, 2016; for a review see Leyton, 1992). Are such properties within the purview of visual impressions, per se, or might they instead reflect considered judgments about what high-level properties subjects *think* should be associated with certain visual cues? And what methods can researchers use to find out?

Perception vs. Judgment

Previous research has emphasized that the apprehension of properties such as causality and animacy sometimes shares many (or even all) of the key features of visual processing — e.g. being driven (even unconsciously) by subtle display details (e.g. Gao et al., 2009; Moors et al., 2017); operating largely regardless of one's beliefs or intentions (e.g. Scholl & Gao, 2013; Schlottmann & Shanks, 1992); influencing the perception of other clearly visual properties (e.g. Scholl & Nakayama, 2004); emerging early in life (e.g. Csibra, 2008; Newman et al., 2008); manifesting in a highly-consistent way across cultures (e.g. Barrett et al., 2005); and operating in a fast (and often irresistible) way in all individuals, excepting those with particular neuropsychological impairments (e.g. Heberlein & Adolphs, 2004). Moreover, some such processing exhibits properties consistent *only* with visual processing, such as retinotopically specific adaptation (e.g. Kominsky & Scholl, 2020; Rolfs et al., 2013).

These findings suggest that we can truly *perceive* such higher-level properties — but of course we can and do *think about* them as well. So how are we to decide whether the results of any specific experiment reflect perception or judgment? The most common dependent measures in such experiments, alas, are relatively unhelpful in this regard. By far the most

typical measure in such work involves simply asking subjects (in one way or another, e.g. with ratings or free reports) about what they "see". For example, researchers studying perceived animacy might ask observers to rate displays on a scale from "definitely not seen as alive" to "definitely seen as alive". The problem with such measures is that words like "see" have many meanings, only some of which implicate visual processing, per se (Dretske, 1969).

Suppose you and two friends are at an art gallery looking at a painting, and you ask them to describe what they see. Your first friend (who has a penchant for stating the obvious) says: "I see red". Your second friend (who was a humanities major) says: "I see poetic justice". Both use the same word — "see" — but in importantly different senses. While redness is a property of one's visual phenomenology, one cannot directly *see* poetic justice in the same way; rather, one *infers it on the basis of* visual information. Or, suppose you run an experiment in which subjects rate moving shapes in terms of how alive they look — and suppose subjects give higher ratings to shapes that dramatically change speeds and/or headings. Do these ratings reflect subjects' percepts, or merely their *judgments* about what stimulus features likely connote animacy? In other words, when an observer in such an experiment says (via their ratings) "I see life", is this more like seeing red or more like seeing poetic justice?[1]

The Current Project: The Blindfold Test

Here we introduce a method that can help to decide between perceptual vs. judgmental interpretations of such results under particular circumstances — the 'Blindfold Test'. While this test is not a magic bullet — and in most cases cannot render any definitive verdict about such questions — it can nevertheless be an important tool. In particular, 'failing' this test effectively identifies cases in which the fact that subjects attest in some way to seeing a property cannot (and should not!) be taken as evidence for visual processing of that property. And conversely,

[1] In this brief empirical report, we cannot defend the distinction between perception and higher-level judgment more generally, though this has been done vigorously elsewhere (e.g. Block, 2023; Firestone & Scholl, 2016).

'passing' this test identifies cases in which such testimony can be trusted as reflecting percepts rather than considered judgments.

The essence of the Blindfold Test is just this: if the results of an experiment with visual stimuli can be replicated when simply *describing* the relevant displays — i.e. *without any visual stimuli at all* (as if the observers were wearing blindfolds) — then the results should not be taken to necessarily implicate visual processing; the jury is still out. In effect, failing the test in this way highlights cases where higher-level (non-visual) judgment can be readily mistaken for perception because *judgments would yield identical results* — and so researchers should be especially cautious when interpreting such results in terms of specific mental processes. And by the same token, passing the test — such that responses with visual stimuli vs. descriptions yield very different patterns of results — can be taken as evidence that the original experiments did *not* simply reflect higher-level reasoning.

Here we report three case studies of the Blindfold Test in action — with two studies that fail the test, and a third that passes. In Experiments 1a and 1b, we explore the apprehension of animacy from motion. In Experiments 2a and 2b, we explore the apprehension of physical forces involved in launching and shattering. And in Experiments 3a and 3b, we explore visual awareness in the first place, in studies of motion-induced blindness for stimuli varying in brightness.

Experiments 1a and 1b: Ratings of Animacy from (Described) Motion

In a first application of the Blindfold Test, we replicated a study from the literature on perceived animacy from motion. Subjects in the original experiment viewed single shapes, which moved along a linear path and then changed speed and/or heading (Figure 1; Tremoulet & Feldman, 2000). Shapes that underwent larger heading and/or speed changes were rated as

more alive (on a 1-7 scale). Do these results, which we replicate here (without visual displays), reflect *perceived* animacy, or might they instead reflect higher-level judgment?

Method

Participants. 100 subjects were recruited online through Amazon's Mechanical Turk and were compensated for their participation with a small monetary payment. We worried that online subjects might be less reliable than in-lab subjects and so we decided before data collection began to use a sample size more than four times that of the original study (and this same sample size was subsequently fixed to be identical in all six experiments reported here).

Materials and Procedure. Subjects first read the following prompt, based on the instructions used in the original experiment: "Pretend you have just been hired as a laboratory technician, and that your job is to determine which of a set of microscopic particles is alive and which is not. Read the following descriptions of particles' movements, and indicate how alive you think each one is on a scale from 1 (definitely not alive) to 7 (definitely alive). Please give a low rating to any particle whose motion seems artificial, mechanical, or strange." Subjects then used Likert scales to rate the movements of described particles. Descriptions began with "*A particle enters the viewing area and moves at a constant speed for 375ms*", followed by a sentence describing a direction change (no change, or turning 10°, 20°, 40° or 80° to its right) and a speed change (to one half, one, two or four times the original speed). Here are three examples:

1. *A particle enters the viewing area and moves at a constant speed for 375ms. It then turns 10 degrees to its right and moves at twice the speed for an additional 375ms.*
2. *A particle enters the viewing area and moves at a constant speed for 375ms. It then continues to move in the same direction, at half the speed, for an additional 375ms.*
3. *A particle enters the viewing area and moves at a constant speed for 375ms. It then continues to move in the same direction, at the same speed, for an additional 375ms.*

Subjects rated the 20 possible descriptions (by clicking radio dials), which were presented on a single page in a randomized order.

Results

Fourteen subjects provided the very same rating for every description. We suspect that these subjects were responding in a principled (rather than a lazy) way, as they reported details about the study and gave reasoned explanations of their decision strategies in response to debriefing questions at the end of the experiment. (For example, one who responded with all 2's wrote: "I thought the movement of the particles … could have been explained in each case by changes in heat.") Data were analyzed in the most conservative way possible, including these subjects. (If they are removed, the results remain the same, except that the already-significant effects become even stronger.)

The effect of direction change on animacy ratings in the original study with visual stimuli (Tremoulet & Feldman, 2000) is depicted in Figure 2a, and the corresponding data from the present experiment with only written descriptions are depicted in Figure 2b. Inspection of these panels suggests that the current results were qualitatively identical to those of the original study — such that greater direction changes led to higher animacy ratings. This impression was verified by the following test. Ratings were submitted to a 5 (direction change: 0°, 10°, 20°, 40°, 80°) x 4 (speed change: .5, 1, 2, 4) repeated measures ANOVA, with Huynh-Feldt corrections applied whenever Mauchly's Test detected a violation of sphericity. This revealed a main effect of direction change, $F(4,396)/(2.11,208.49)=14.58$, $p<.001$, $\eta p^2=.13$.

The effect of speed change on animacy ratings in the original study is depicted in Figure 2d, and the corresponding data from the present experiment are depicted in Figure 2e. Inspection of these panels similarly suggests a qualitative replication, and indeed: as in the original study, greater final speeds were associated with higher animacy ratings, $F(3,297)/(1.77,174.83)=9.71$, $p<.001$, $\eta p^2=.09$.

Tremoulet and Feldman (2000) also report an interaction, whereby increasing direction changes led to greater increases in animacy ratings for particles with lower final speeds than for particles with higher final speeds. In the present experiment, however, no such interaction was observed, $F(12,1188)/(9.67,957.75)=1.02$, $p=.423$, $\eta p^2=.01$.

Direct Replication

Given the ease of running online studies, we directly replicated this study with a new sample of 100 subjects in Experiment 1b.

Thirteen subjects provided the very same rating for every description, and these were again retained in our analyses. (Again, if these subjects are removed, the results remain the same, except that the significant effects become even stronger.)

The effect of direction change on animacy ratings in Experiment 1b is depicted in Figure 2c. Inspection of this figure suggests that the current results were again qualitatively identical to those of the original study (and to those of Experiment 1a): again, greater direction changes elicited higher animacy ratings. A 5x4 repeated measures ANOVA confirmed this main effect of direction change, $F(4,396)/(3.17,314)=20.69$, $p<.001$, $\eta p^2=.17$.

The effect of speed changes on animacy ratings in Experiment 1b is depicted in Figure 2f. Inspection of this figure suggests that, here too, the results were qualitatively identical to those of the original study (and to those of Experiment 1a): again, greater final speeds were associated with higher animacy ratings, $F(3,297)/(2.17,214.94)=30.32$, $p<.001$, $\eta p^2=.23$.

Unlike Experiment 1a, however, this replication also observed Tremoulet and Feldman's (2000) interaction between heading and speed changes, $F(12,1188)/(10.28,1018.11)=2.49$, $p=.006$, $\eta p^2=.02$. (While the original paper reports no statistical tests characterizing the observed interaction, the present results resemble the qualitative description provided. As in the original study, direction changes influenced animacy ratings more for displays with lower final speeds than for displays with higher final speeds.)

Discussion

Does the effect of speed/direction changes on reports of greater animacy reflect perceptual impressions, or higher-level judgments? The original report of this effect (Tremoulet & Feldman, 2000) does not take a strong stand on this question, one way or the other. On one hand, they sometimes interpret such results in terms of the subjects attempting to "classify" (p. 943), "decide" (p. 944), or "judge" (p. 946) whether the dot was animate — and they are careful to note that "it is not clear to what extent these effects are truly perceptual" (p. 950). But on the other hand, they also take themselves to be studying "the subjective impression of animacy" (p. 944), as computed by "the human visual system" (p. 943) in "a relatively immediate and 'bottom-up'" manner (p. 947). And many subsequent reports have also unambiguously interpreted these results in terms of visual perception, per se — e.g. referring to "a perceptual phenomenon which reflects visual processing" (Di Giorgio et al., 2021, p. 1), or "bottom-up processes, such as visual cues to animacy" (Gerstenberg & Tenenbaum, 2016, p. 40).

The present study replicated the effect of speed/direction changes on reports of greater animacy, even though the visual displays were replaced with written descriptions. And as a result, the original study fails the Blindfold Test: since similar results can arise via higher-level judgments (without subjective impressions), the original results may also reflect such judgments, and needn't appeal to perception at all. (And both sets of results could readily reflect task demands, since as a subject there is pressure to vary one's responses across trials, and to do so based on speed and direction changes, which are so clearly properties which are being manipulated.) Of course, these results do not entail that the original finding *must* reflect higher-level reasoning, but they do demonstrate how the original finding alone fails to implicate visual perception, per se.

Experiments 2a and 2b: Ratings of Force in (Described) Launching and Shattering

A strength of the Blindfold Test is that it can be applied across many types of putatively perceptual effects. And so to demonstrate this generality, we also applied it to a study from another domain in high-level vision — involving perceived physical forces. In such experiments, subjects typically rate "launchers" as exerting more force on their "targets" than vice versa (e.g. White, 2007). One recent study, however, found that this asymmetry in force ratings is flexible, depending on how the launcher and target behave after contact (Hubbard & Ruppel, 2013). Subjects viewed launching events where one or both shapes 'shattered' into four or nine pieces upon contact (Figure 3), and they rated the amount of force imparted by the launcher to the target, and vice versa. Launchers were rated as exerting more force than targets when the target shattered and the launcher remained intact, and vice versa (Figure 4a). Might this result also reflect higher-level judgments rather than visual impressions?

Method

Participants. 100 subjects were recruited online through Amazon's Mechanical Turk and were compensated for their participation with a small monetary payment. This sample size (which was more than five times that used in the original study) was chosen to match that of Experiments 1a and 1b.

Materials and Procedure. The design of the experiment mirrored that of Experiment 2 in Hubbard and Ruppel (2013), but with the visual displays replaced by written descriptions of eight possible shattering events. Descriptions always began with a red object moving toward a blue object, and ended with one or both shattering into four or nine pieces as soon as the objects touched (e.g. "A red object moves toward a blue object. As soon as the red object touches the blue object, the red object remains intact and the blue object breaks into four pieces."). In the T4 and T9 events, the launcher remained intact and the target shattered into either four or nine pieces. In the L4 and L9 events, the target remained intact and the launcher shattered into

either four or nine pieces. In the T4L4, T4L9, T9L4, and T9L9 events, the launcher and target both shattered into four or nine pieces.

Each subject completed two blocks of ratings. In the "launcher ratings" block, they typed numbers from 0 (no force at all) to 100 (maximum possible force) to rate the amount of force that the red object exerted on the blue object in each event. In the "target ratings" block, they rated the force that the blue object exerted on the red object in each event. Within both blocks, descriptions were presented on a single page in a randomized order. The order of the blocks was counterbalanced between subjects.

Results

As in the original study, force ratings were analyzed in a 2 (source: launcher, target) × 8 (event type: T4, T9, L4, L9, T4L4, T4L9, T9L4, T9L9) repeated measures ANOVA, with Huynh-Feldt corrections applied whenever Mauchly's Test detected a violation of sphericity. The results mirrored those of the original experiment. There was as main effect of source, $F(1,99)=11.34$, $p=.001$, $\eta p^2=.10$, with the launcher rated as exerting more force (95% CI [55.73, 62.24]) than the target (95% CI [47.44, 55.24]); a main effect of event type, $F(7,693)/(4.99,494.16)=31.90$, $p<.001$, $\eta p^2=.24$; and an interaction, $F(7,693)/(2.53,250.30)=23.61$, $p<.001$, $\eta p^2=.19$. The original paper did not draw its main conclusions from these statistics, but rather from informal observations of how event types differed in terms of the relative force judged to be exerted by the launcher and target. We turn to these now.

Force Ratings when One Object Shattered. The original study's results for conditions in which only one object shattered are depicted on the left side of Figure 4a. As can be seen from the left side of the figure, the strongest evidence for a flexible asymmetry in force ratings comes from the four conditions in which one object shattered and the other remained intact. If the target shattered and the launcher did not shatter (T4, T9), the launcher was rated as exerting more force than the target. If the launcher shattered and the target did not shatter (L4, L9), the

target was rated as exerting more force than the launcher. The results of the present study for these conditions are depicted on the left side of Figure 4b, and inspection of this panel indicates that these trends replicated robustly. (The original study did not report any analyses of these patterns, but in our replication this pattern was statistically robust: for example, the signed difference in force ratings between the launcher and the target was greater for the T4/T9 conditions than for the L4/L9 conditions, $t(99)=6.13$, $p<.001$, $d_z=.61$.)

Force Ratings when Both Objects Shattered. The right sides of Figures 4a and 4b depict force ratings in conditions where both objects shattered. Here again, the results of the present study strongly resemble those of the original. When the target shattered into four pieces and the launcher also shattered (T4L4, T4L9), force ratings were similar. When the target shattered into nine pieces and the launcher also shattered (T9L4, T9L9), the launcher was rated as exerting more force than the target.

Figure 5a extracts the most robust and meaningful trends in Hubbard and Ruppel (2013) in terms of *differences* in force ratings between the launcher and the target. Across all event types, the launcher was rated as exerting more force than the target. However, there was a flexible asymmetry in force ratings, which was clearest in conditions where only one object shattered. If the target shattered and the launcher did not (T4, T9), the launcher was rated as exerting more force than the target. But if the launcher shattered and the target did not (L4, L9), then the target was rated as exerting more force than the launcher. The results of the present study are depicted in Figure 5b, and inspection of this panel indicates that these key trends all replicated. (The original study did not report any comparisons of these effects, but in our study, each bar in this graph depicts a difference score that was significantly different from both of the others, all *p*s < .001.)

Direct Replication

Given the ease of running online studies, we directly replicated this study with a new sample of 100 subjects in Experiment 2b.

There was again a main effect of source, $F(1,99)=18.87$, $p<.001$, $\eta p^2=.16$, with the launcher rated as exerting more force (95% CI [54.04, 60.03]) than the target (95% CI [42.24, 50.41]); a main effect of event type, $F(7,693)/(5.82,576.49)=24.12$, $p<.001$, $\eta p^2=.20$; and an interaction, $F(7,693)/(2.50,247.80)=26.64$, $p<.001$, $\eta p^2=.21$.

Force Ratings when One Object Shattered. The results of Experiment 2b are depicted in Figure 4c. Inspection of the left side of this figure indicates that the results were qualitatively identical to both the original study and Experiment 2a. If the target shattered and the launcher did not shatter (T4, T9), the launcher was rated as exerting more force than the target. But if the launcher shattered and the target did not shatter (L4, L9), the target was rated as exerting more force than the launcher. And again, the signed difference in force ratings between these groups was reliable, $t(99)=6.49$, $p<.001$, $d_z=.65$.

Force Ratings when Both Objects Shattered. The right side of Figure 4c depicts force ratings in conditions where both objects shattered. As can be appreciated from these graphs, these results resemble those of both the original study and Experiment 2a. When the target shattered into four pieces and the launcher also shattered (T4L4, T4L9), force ratings were roughly equal. But when the target shattered into nine pieces and the launcher also shattered (T9L4, T9L9), the launcher was rated as exerting more force than the target.

Figure 5c depicts the key trends from these results, which again qualitatively replicated the patterns for both the original study (Figure 5a) and Experiment 2a (Figure 5b) — again with each bar in this graph depicting a difference score that was significantly different from both of the others, all ps < .001.

Discussion

Does the flexibility of force ratings for launchers and targets reflect perceptual impressions, or higher-level judgments? The original report of this effect (Hubbard & Ruppel, 2013) does not focus directly on this issue, and mostly discusses the results in operationalized terms (focusing on the "ratings"). At the same time, however, these results are taken to bear on

"theories of phenomenal causality" (p. 987), and to reflect "perception of force and perception of causality" (p. 1004), as driven by "a pattern of visual stimulation" (p. 1007) — and this result has been given similarly perceptual interpretations by others, e.g. referring to "visually perceived events" (Vicovaro et al., 2023 p. 2), or "signature perceptual features" (Danks & Dinh, 2022, p. 82).

The present study, however, shows that no visual stimulation is required after all, due to a failure in the Blindfold Test: the current results replicated the effect of post-impact shattering on reports of perceived force, even though the visual displays were replaced with written descriptions. And so once again, these results do not entail that the original finding *must* reflect higher-level reasoning, but they do demonstrate how the original finding alone fails to implicate visual processing.

Experiments 3a and 3b:
Motion-Induced Blindness as a Function of (Described) Target Luminance

As a case study of an effect that seemed likely to *pass* the Blindfold Test, we next turned to a phenomenon of visual awareness that (like the displays in Experiments 1 and 2) seems especially easy to describe: *Motion-Induced Blindness* (MIB; e.g. Bonneh et al., 2001; Graf et al., 2002; New & Scholl, 2008, 2018). A typical MIB experiment contains a salient global motion pattern (such as a rotating blue background texture) presented along with static elements (e.g. salient yellow discs). When subjects fixate on such displays, they reliably experience a dramatic illusion in which the salient static elements disappear from visual awareness — often for several seconds at a time, and even when fully attended (Bonneh et al., 2001; Schölvinck & Rees, 2009)! It has been argued that this blindness does not actually reflect any sort of failure or limitation of the visual system; instead, MIB may reflect an adaptive unconscious inference, as the visual system actively erases the static elements from awareness, for the same reason you do not see

the shadows from the blood vessels that line the retina (New & Scholl, 2008, 2018): both remain retinotopically stable, and fail to "play along" with surrounding dynamic events. In effect, MIB may be a case in which the visual system interprets a bit of stimulation not as a part of the external world, but as an artifact of itself.

In the present context, we focused on a particular foundational result related to the conditions that give rise to MIB: against a dark background, brighter static stimuli are *more* likely to disappear than are darker stimuli (Bonneh et al., 2001). Does this effect reflect brute visual impressions, or higher-level inferences about what conditions are likely to make stimuli disappear during MIB?

Method

Participants. 100 subjects were recruited online through Amazon's Mechanical Turk and were compensated for their participation with a small monetary payment. This sample size (which was ten times that used in the original study) was chosen to match that of Experiments 1a, 1b, 2a, and 2b.

Materials and Procedure. The design of the experiment mirrored the experiment depicted in Figure 2a of Bonneh et al. (2001), but with the visual displays replaced by written descriptions. Subjects first read a few sentences describing the experimental setup:

> Some researchers have invited you to participate in an experiment on visual perception. You view a display that contains 150 small blue dots clustered together in a circular region, on a black background. These dots move together for 60 seconds, as if they are stuck on the surface of an invisible, rotating sphere. Centered in front of the "sphere" are three larger yellow discs. It turns out (strangely enough) that when you stare at the center of this kind of display, sometimes the yellow discs seem to disappear for a few seconds.

Subjects then read four descriptions in a randomized order, one for each of the display conditions in the original experiment: "The yellow discs are shown at [10/20/40/80]% brightness." For each of these, they were asked "What percent of the time do you think that any of the discs would disappear?", and they used a slider to provide a rating from 0-100.

Results

The results of the original experiment from Bonneh et al. (2001) are depicted in Figure 6a, and the current results are depicted in Figure 6b. As is immediately apparent, these two experiments produced opposite patterns of results: the original study (with visual stimuli) found that increasing the brightness of the discs *increased* MIB rates, whereas our subjects predicted the exact opposite — that brighter discs would seem to disappear *less* often (all t's>6.31, all p's<.001, all d_z's>.63).

Direct Replication

Given the ease of running online studies, we directly replicated this study with a new sample of 100 subjects in Experiment 3b. As can be seen in Figure 6c, subjects again predicted that brighter discs would disappear less often (all t's>6.80, all p's<.001, all d_z's>.68).

Discussion

In contrast to Experiments 1 and 2, Experiment 3 *passed* the Blindfold Test: when the original displays were replaced with written descriptions, subjects' judgments about how these displays should look diverged dramatically from what they would have said when experiencing them directly. This constitutes compelling evidence that the original experiment did *not* simply reflect higher-level reasoning about what would be likely to happen, but rather reflected actual visual impressions.

General Discussion

In the study of visual cognition, a common experimental strategy involves manipulating visual displays, and then assessing subjects' impressions of those displays by directly asking them about the results of those manipulations — e.g. "Did *that* event look causal?", or "What animacy rating would you give to *those* objects?", or "Did the targets disappear?" In some cases, such responses may truly reflect visual impressions, but in other cases they may instead arise

due to the often-inconvenient fact that subjects can also *think about* what effects manipulations *should* have. When should we take subjective reports as evidence of what subjects *see*? The present study's answer is: only when they pass the Blindfold Test — such that the same responses do not result from mere descriptions of the displays, without any actual visual stimulation.

In essence, the Blindfold Test empirically identifies a subtle but critical confound — wherein the results that are obtained from seeing a scene are confounded with those that would be obtained from mere descriptions of the scene.[2] Whenever this is the case, you shouldn't conclude that an effect is truly perceptual without some additional evidence for that conclusion, because you've identified a situation wherein demonstrably non-perceptual judgments will yield the identical outcome. Harkening back to the art gallery, the Blindfold Test flags experimental results that may implicate "seeing" more in the sense of seeing poetic justice than in the sense of seeing redness. This logic applies to the results of both Experiments 1 and 2. As such, absent any other evidence, we don't think these studies necessarily implicate visual perception: their results may reflect only subjects' higher-level reasoning about how animate entities are likely to behave (in Tremoulet & Feldman, 2000), or how shattering is related to physical force (in Hubbard & Ruppel, 2013).

In contrast, when a study passes, the Blindfold Test provides compelling evidence *against* the possibility that the initial results with visual stimuli reflected higher-level reasoning about the displays. From Experiment 3, for example, we can be confident that the original result — that brighter targets disappear more during MIB (Bonneh et al., 2001) — reflects actual visual experience rather than reasoning, since subjects reason that the opposite pattern is likely to hold.

[2] In this sense, the test is a sort of 'overgeneralization' measure (Firestone & Scholl, 2015), in that it focuses on cases in which the same experimental design overgeneralizes to unambiguously non-perceptual cases.

Relationships to Past Work

To our knowledge, nobody has previously formalized a test like this as such — but there are several related precedents. First, in the literature on so-called 'embodied' perception, researchers have questioned whether certain effects (e.g. in which subjects wearing a heavy backpack report that hills look steeper) reflect changes in subjects' perception, or rather their reasoning about what they *ought* to see, given the task demands of the experimental situation. Here a useful strategy for deciding has been to administer a post-experiment debriefing survey, asking subjects what effect they *predicted* the heavy backpack should have on their perception. In these replications, the effect of wearing a heavy backpack on perceived hill slant has turned out to be driven entirely by those subjects who guess the hypothesis — suggesting that reasoning due to task demands, rather than an effect of wearing a heavy backpack on perception, most likely causes the effect (Durgin et al., 2009; 2012).

Second, recent work on the 'rubber hand illusion' has focused on subjects who are shown a video of the experiment setup, and are then asked to predict the perceptual effects of synchronized visual-tactile signals. Such subjects correctly predict the results of the study, raising the prospect that such task-demands explain the results even during the live 'perceptual' conditions (Lush, 2020).

Third, work in intuitive physics has often contrasted the apprehension of physical forces and masses from dynamic visual displays vs. static diagrams and verbal descriptions, and has taken differences in these different kinds of judgments as evidence for distinct underlying processing (e.g. Kaiser et al., 1985; for a review, see Vicovaro, 2023). The contribution of the present paper is thus to formalize the logic of this sort of test, and to demonstrate how it can be applied broadly to aid the interpretation of a variety of putatively perceptual effects.

Implementing the Blindfold Test in Practice

We have seen that the Blindfold Test can constrain interpretations of the mental processes underlying certain experimental results, in terms of whether their results can be taken

as reflecting the operation of visual perception, per se — where the test can both question and support this possibility. But it is also important to stress at least three limitations of this test:

First, note that studies which pass the Blindfold Test are likely to be those whose manipulations are relatively counterintuitive — a property which will not always hold, and which sometimes may not be possible. The first two case studies in the present project (in Experiments 1 and 2), for example, failed the Blindfold Test for what seems like a straightforward reason: in each case, the function relating the key variable to the resulting performance seemed especially intuitive. It just makes sense that an object which has a greater sudden change in velocity or direction is more likely to be animate, for example. Such reasoning may or may not be effectively implemented as a kind of automatic "unconscious inference" during visual processing itself, but it is *surely* implemented as a form of higher-level thought. In contrast, the key result of Experiment 3 involved a strikingly counterintuitive result: it just seems so unlikely that more salient objects would be *more* likely to disappear from awareness! As such, the Blindfold Test helps to catch studies of highly intuitive manipulations — while also emphasizing an important theoretical advantage of counterintuitive effects, when one is interested in identifying underlying mental mechanisms.

Second, the Blindfold Test may only be reliable for studies whose visual displays are readily put into words. This was true for all three of the current case studies, and indeed part of their appeal is how simple and straightforward such displays are. Describing these dynamic scenes (and their underlying manipulations) is thus relatively easy, requiring relatively few words and a relatively low memory load. But when this is not the case, a study might seem to pass the Blindfold Test only because the descriptions are insufficient for subjects to understand what the displays are actually like — perhaps because of subtle nuances (which are too difficult to describe), or perhaps due to brute complexity (when descriptions may be too long and cumbersome for comprehension). Such complexity is sometimes unavoidable, when one is interested in subtle displays with relatively ineffable qualities — e.g. involving nuanced facial

expressions (e.g. Todorov, 2017), or complex dynamic patterns that would require hundreds of words to fully describe, despite seemingly simple percepts (such as the dynamics of fluids [Kawabe et al., 2015], fine particles [vanMarle & Scholl, 2003], or soft materials [Wong et al., 2023]). If an experiment passes the Blindfold Test, one must thus consider on a case-by-case basis whether it passed just due to incomplete or misleading descriptions. But for experiments with sufficiently simple displays, one will quickly exhaust the list of additional details which one could additionally provide for subjects to make different responses.

While accepting the in-principle limitation noted in the previous paragraph, it is also possible to combat it in practice. In particular, the Blindfold Test may be rendered applicable to experiments with more complex displays by first showing subjects a starting 'baseline' display (either in full, or as a diagram), and then simply describing simple manipulations of that already-perceived display. But the details of this approach (what display to show, what to verbally describe, and which judgments to leave to the subject to make) will depend greatly on the specific putative visual processing that is under study. In general, there is a question of how one should describe a study's displays (e.g. in terms of pixels vs. terms of moving objects). It seems fine to provide a description which does some of the work which is thought to be done by visual processing, to target the original study's claim about how such perceptual interpretations drive a further perceptual interpretation. (And replacing dynamic displays with static diagrams and descriptions of events may also be helpful for a targeted test of whether the perception of objects' movements, per se, was necessary for an effect to occur.) The key principle here is that implementations of the Blindfold Test should not directly visually depict or display the *relevant* property, whose 'perceptual' status is at issue — even if some of the other variables or properties are visually presented.

Third, and most generally, it seems important to emphasize a foundational asymmetry in the conclusions that the test can support: when a study passes the Blindfold Test, this may constitute clear evidence *against* explanations that appeal only to higher-level reasoning. But

when a study fails the Blindfold Test, that does not constitute evidence *for* explanations that appeal only to higher-level reasoning. Rather, such failures license only negative inferences — that the original results cannot by themselves be taken as evidence in favor of perceptual explanations. In such circumstances, the jury is thus still out, and other evidence would be required to test whether the results reflect visual processing, per se. What forms could that evidence take?

Surviving the Blindfold Test

We see (!) at least three broad possibilities for different kinds of evidence that could still support a perceptual interpretation of an effect that fails the Blindfold Test:

First, researchers can seek manipulations that elicit not only different patterns of subjective reports, but also robust differences in shared visual phenomenology. In fact, the absence of just this sort of vivid phenomenology was what led us to initially suspect that the effects of speed and direction on a single dot's movement (in Tremoulet and Feldman, 2000) might not reflect perception after all: these displays simply don't *look* so vividly animate in the way that has inspired this field ever since the seminal studies of Heider and Michotte. In arbitrating between perceptual and cognitive interpretations, however, such phenomenology can be definitive: would you ever wonder, for example, whether illusory contours (as when some carefully placed 'pac-men' cause you to perceive an illusory triangle; e.g. Kanizsa, 1955) might only reflect what subjects *think* should be in the display, with no corresponding percept? Never — simply because such percepts are so vivid and unmistakable.

Second, researchers can supplement subjective reports with objective performance measures that exploit the *limits* of higher-level thought in order to study perception. In other recent studies, for example, subjects' ability to detect animate 'chasing' in dynamic visual displays follows precise psychophysical functions, such that subjects are unable to simply *decide* to treat some stimulus or another as reflecting chasing, even when they have every incentive to do so, and when they know that chasing actually exists in those cases (Gao et al., 2009; Gao &

Scholl, 2011; van Buren et al., 2017). Such performance measures, rather than mimicking patterns of judgment, illustrate how perception and judgment can conflict.

Finally, in cases where a study fails the Blindfold Test, its conclusions about perception could still be bolstered by appeal to properties of the displays that can be explained *only* by visual processing. In the perception of causality, for example, such displays can yield retinotopically specific patterns of visual adaptation — which no 'judgmental' account could ever hope to explain (Kominsky & Scholl, 2020; Rolfs et al., 2013).

Conclusion: Identifying 'Anti-Illusions'

The study of visual perception is suffused with visual illusions, and the essence of an excellent illusion is that it persists despite certain knowledge that one's perception does not match reality. In other words, at the root of most visual illusions is a stark conflict between perception and judgment (van Buren & Scholl, 2018). In this sense, the Blindfold Test identifies "anti-illusions" in the study of visual cognition — cases where perception and judgment are in exact *alignment*. And, when such anti-illusions are present, psychologists should be cautious about implicating visual processing based on subjective reports about visual displays.

References

Barrett, H., Todd, P., Miller, F., & Blythe, M. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution & Human Behavior, 26*, 313-331.

Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.

Bonneh, Y. S., Cooperman, A., & Sagi, D. (2001). Motion-induced blindness in normal observers. *Nature, 411*, 798-801.

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.

Choi, H., & Scholl, B. J. (2006). Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception. *Perception, 35*, 385-399.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition, 107*, 705-717.

Danks, D., & Dinh, P. N. (2022). Causal perception and causal inference: An integrated account. In P. Willemsen & A. Wiegmann (Eds.), *Advances in Experimental Philosophy of Causation* (pp. 81-100). Bloomsbury.

Di Giorgio, E., Lunch, M., Vallortigara, G., & Simion, F. (2021). Newborns' sensitivity to speed changes as a building block for animacy perception. *Scientific Reports, 11*:542, 1-19.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review*, *16*, 964-969.

Durgin, F. H., Ruff, A. J., & Russell, R. (2012). Constant enough: On the kinds of perceptual constancy worth having. In G. Hatfield & S. Allred (Eds.), *Visual Experience: Sensation, Cognition and Constancy* (pp. 87-102). Oxford University Press.

Dretske, F. I. (1969). *Seeing and knowing*. Chicago: University of Chicago Press.

Firestone, C., & Scholl, B. J. (2015). When do ratings implicate perception versus judgment? The "overgeneralization test" for top-down effects. *Visual Cognition, 23*, 1217-1226.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral & Brain Sciences, e229,* 1-77.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception & Performance, 37,* 669-684.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology, 59*, 154-179.

Gerstenberg, T., & Tenenbaum, J. (2017). Intuitive theories. In M. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 515-548). Oxford University Press.

Graf, E. W., Adams, W. J., Lages, M. (2002). Modulating motion-induced blindness with depth ordering and surface completion. *Vision Research, 42*, 2731-2735.

Heberlein, A. S., & Adolphs, R. (2004). Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences, 101*, 7487-7491.

Heider F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243-259.

Hubbard, T. L., & Ruppel, S. E. (2013). Ratings of causality and force in launching and shattering. *Visual Cognition, 21,* 987-1009.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 11, 795-803.

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Rivista di Psicologia*, *49*, 7-30.

Kawabe, T., Maruya, K., Fleming, R. W., & Nishida, S. (2015). Seeing liquids from visual motion. *Vision Research, 109*, 125-138.

Kominsky, J., & Scholl, B. J. (2020). Retinotopic adaptation reveals distinct categories of causal perception. *Cognition, 203*:104339, 1-21.

Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA: MIT Press.

Lush, P. (2020). Demand characteristics confound the rubber hand illusion. *Collabra: Psychology, 6*, 1-10.

Michotte, A. (1946/1963). *La perception de la causalité*. Louvain: Institut Supérieur de Philosophie 1946. English translation of updated edition by T. Miles & E. Miles, *The perception of causality*, Basic Books, 1963.

Moors, P., Wagemans, J., & de-Wit, L. (2017). Causal events enter awareness faster than non-causal events. *PeerJ*, 5:e2932.

New, J. J., & Scholl, B. J. (2008). "Perceptual scotomas": A functional account of motion-induced blindness. *Psychological Science, 19*, 653-659.

New, J. J., & Scholl, B. J. (2018). Motion-induced blindness for dynamic targets: Further explorations of the perceptual scotomas hypothesis. *Journal of Vision, 18*, 1-13.

Newman, G. E., Choi, H., Wynn, K., & Scholl, B. J. (2008). The origins of causal perception: Evidence from postdictive processing in infancy. *Cognitive Psychology, 57*, 262-291.

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology, 23*, 250-254.

Schlottmann, A., & Shanks, D. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly Journal of Experimental Psychology, 44A*, 321-342.

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment? In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, & intention* (pp. 197-230). Cambridge, MA: MIT Press.

Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science, 13*, 493-498.

Scholl, B. J., & Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4,* 299-309.

Schölvinck, M., & Rees, G. (2009). Attentional influences on the dynamics of motion-induced blindness. *Journal of Vision*, *9*:38, 1-9.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton, NJ: Princeton University Press.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29*, 943-951.

van Buren, B., & Scholl, B. J. (2018). Visual illusions as a tool for dissociating seeing from thinking: A reply to Braddick (2018). *Perception, 47*, 999-1001.

van Buren, B., Gao, T., & Scholl, B. J. (2017). What are the underlying units of perceived animacy?: Chasing detection is intrinsically object-based. *Psychonomic Bulletin & Review*, *24*, 1604-1610.

vanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects vs. substances. *Psychological Science, 14*, 498-504.

Vicovaro, M. (2023). Grounding intuitive physics in perceptual experience. *Journal of Intelligence, 11*, 1-20.

Vicovaro, M., Brunello, L., & Parovel, G. (2023). The psychophysics of bouncing: Perceptual constraints, physical constraints, animacy, and phenomenal causality. *PLoS ONE, 18*, e0285448.

White, P. A. (2007). Impressions of force in visual perception of collision events: A test of the causal asymmetry hypothesis. *Psychonomic Bulletin & Review, 14*, 647-652.

Wong, K. W., Bi, W., Soltani, A., Yildirim, I., & Scholl, B. J. (2023). Seeing soft materials draped over objects: A case study of intuitive physics in perception, attention, and memory. *Psychological Science, 34*, 111-119.

Declarations

Funding: This project was funded by an NSF Graduate Research Fellowship awarded to BvB, and by ONR MURI #N00014-16-1-2007 awarded to BJS.

Conflicts of interest: The authors have no conflicts of interest to declare.

Ethics approval: All experimental methods and procedures were approved by the Yale University Institutional Review Board.

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

Consent for publication: The authors affirm that human research participants provided informed consent for publication.

Availability of data and materials: Our submission includes the data for all three experiments.

Code availability: Analysis code will be made available upon request.

Authors' contributions: Both authors developed the study concept, contributed to the study design, and wrote the manuscript. Data collection and analysis were performed by B. van Buren.

Open practices statement: Data or materials for the experiments are available upon request. The experiments were not preregistered, but all analysis procedures were decided before data collection began, to match the analyses of the original studies as closely as possible.

Acknowledgements: For helpful conversation and/or comments on previous drafts, we thank Frank Durgin, Guido Hesselmann, Daglar Tanrikulu, and the members of the New School Perception Lab and Yale Perception & Cognition Lab. We are especially indebted to Chaz Firestone for the term 'Blindfold Test'.

Figure Captions

Figure 1. Depiction of the stimuli used in Tremoulet and Feldman (2000). In the relevant conditions, a 'particle' initially moved in a random direction at a constant speed for 375 ms, then changed its speed or direction (or both, or neither), and continued moving for another 375 ms. (Adapted from Tremoulet & Feldman, 2000)

Figure 2. Animacy ratings as a function of the particle's direction change in (A) Tremoulet and Feldman (2000), (B) Experiment 1a, and (C) Experiment 1b — along with animacy ratings as a function of the particle's speed change in (D) Tremoulet and Feldman (2000), (E) Experiment 1a, and (F) Experiment 1b. Error bars represent 95% confidence intervals, subtracting out the shared variance.

Figure 3. Depiction of some of the stimuli used in Experiment 2 of Hubbard and Ruppel (2013). In all conditions, the launcher moved until it was adjacent with the target. The first row depicts the condition (T4) where the target subsequently shattered into four pieces. The second row depicts the condition (L4) where the launcher subsequently shattered into four pieces. The third row depicts the condition (L4T4) where both objects shattered into four pieces. (Adapted from Hubbard & Ruppel, 2013)

Figure 4. Force ratings for the launcher and target for the eight different event types in (A) Hubbard and Ruppel (2013), (B) Experiment 2a, and (C) Experiment 2b. In (A), error bars represent the standard error of the mean. In (B) and (C), error bars represent 95% confidence intervals, subtracting out the shared variance.

Figure 5. Differences in force ratings given to the launcher and the target in three situations: overall; in the conditions where the launcher remained intact and the target shattered (T4, T9); and in the conditions where the launcher shattered and the target remained intact (L4, L9). The differences in these three situations are depicted for (A) Hubbard and Ruppel (2013), (B) Experiment 2a, and (C) Experiment 2b. Error bars represent 95% confidence intervals, subtracting out the shared variance.

Figure 6. Effects of the discs' brightness in (A) Motion-Induced Blindness (MIB) rates in Bonneh et al. (2001), and predicted MIB rates in (B) Experiment 3a, and (C) Experiment 3b. Error bars in B and C represent 95% confidence intervals, subtracting out the shared variance.

Figure 1

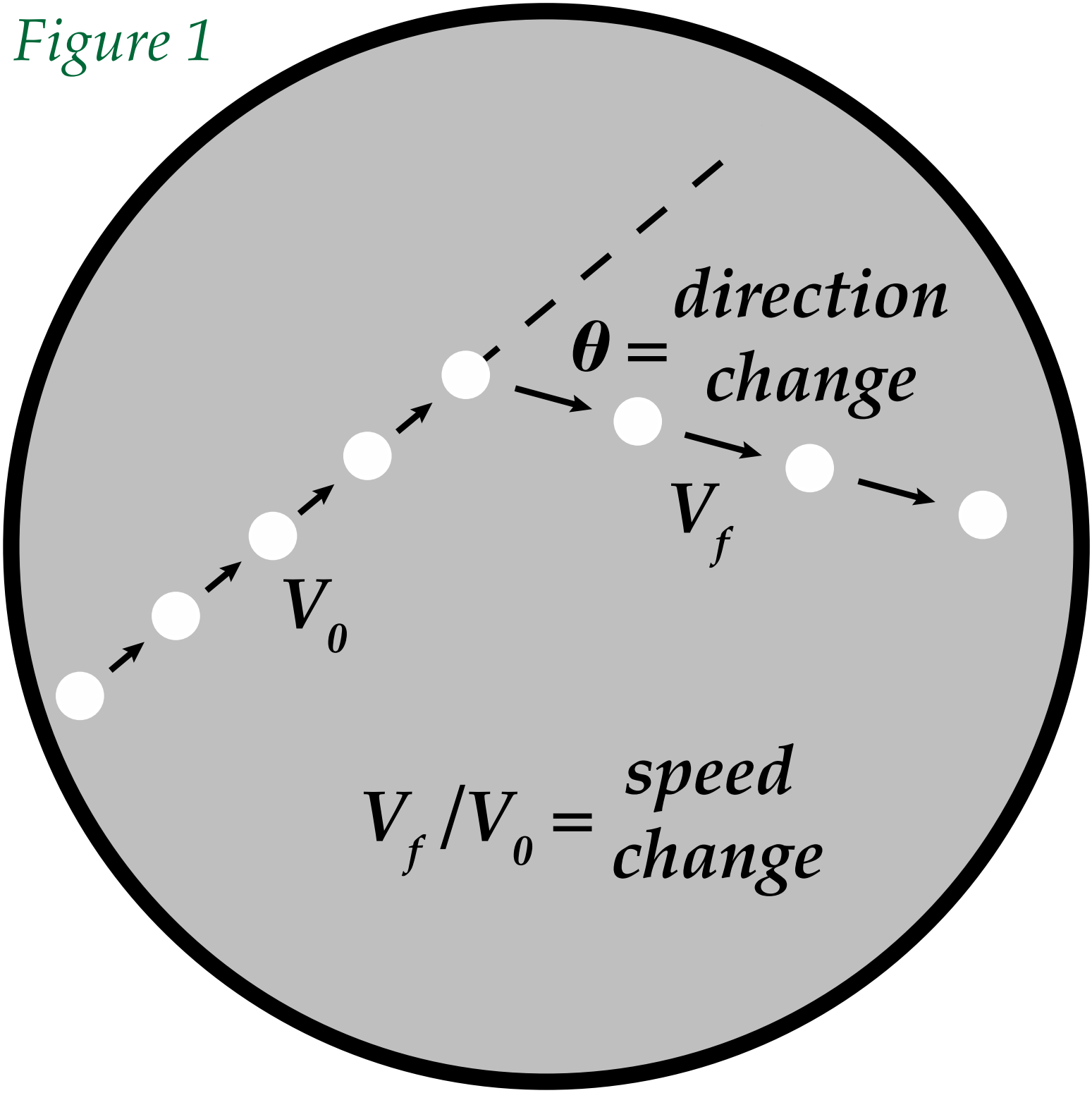


Figure 2

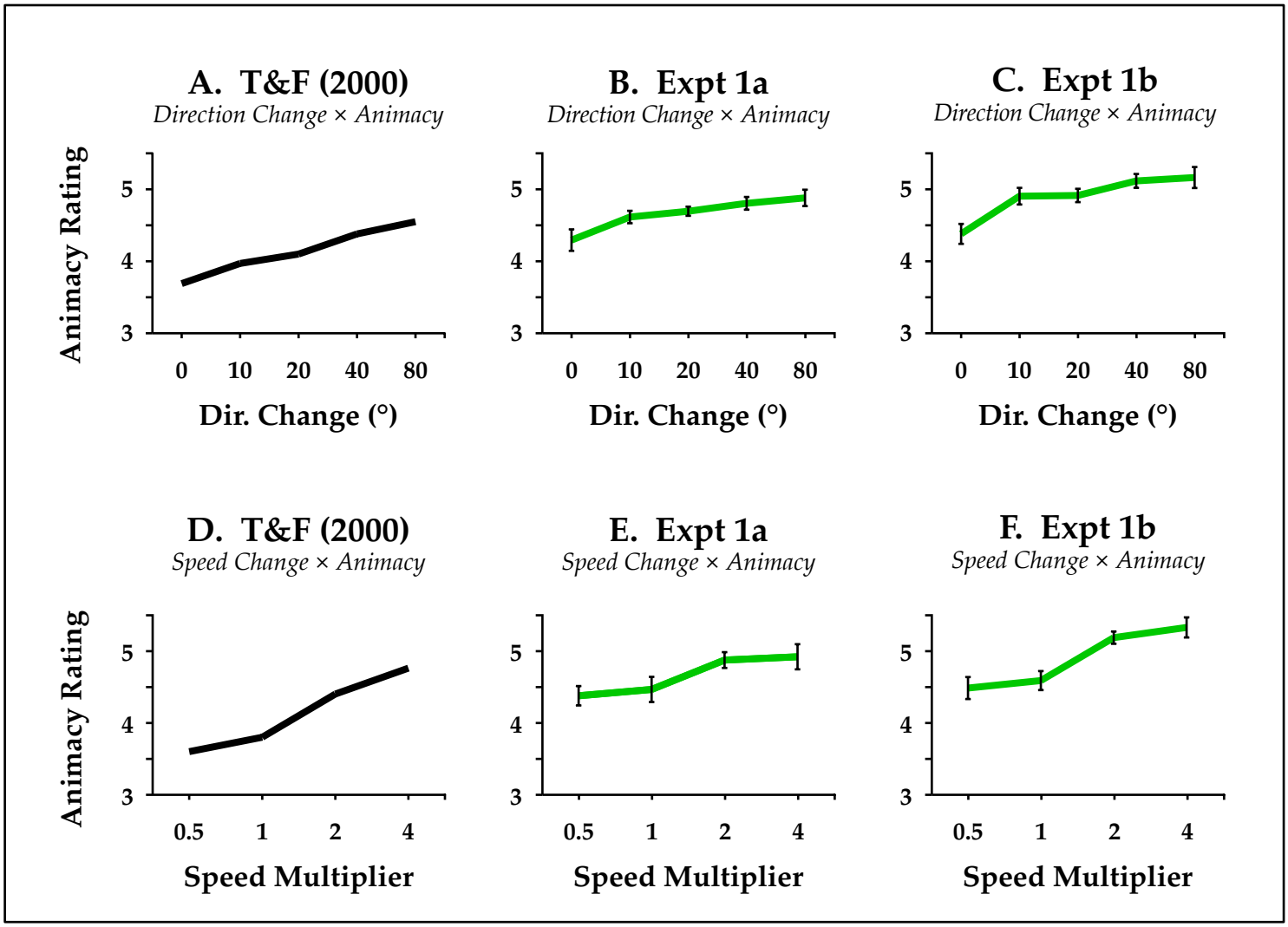

A. T&F (2000)
Direction Change × Animacy
B. Expt 1a
Direction Change × Animacy
C. Expt 1b
Direction Change × Animacy
Animacy Rating
5
4
3
0
10
20
40
80
Dir. Change (°)
D. T&F (2000)
Speed Change × Animacy
E. Expt 1a
Speed Change × Animacy
F. Expt 1b
Speed Change × Animacy
0.5
1
2
4
Speed Multiplier

Figure 3

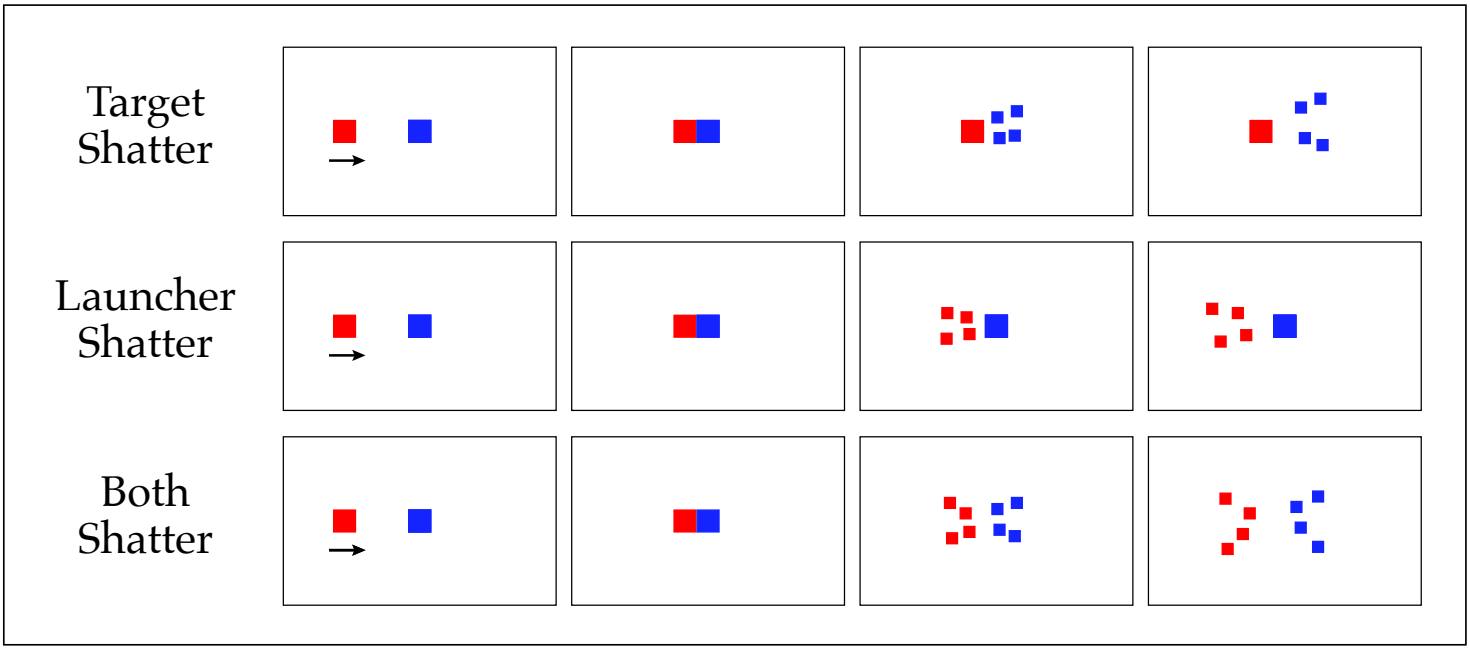


Figure 4

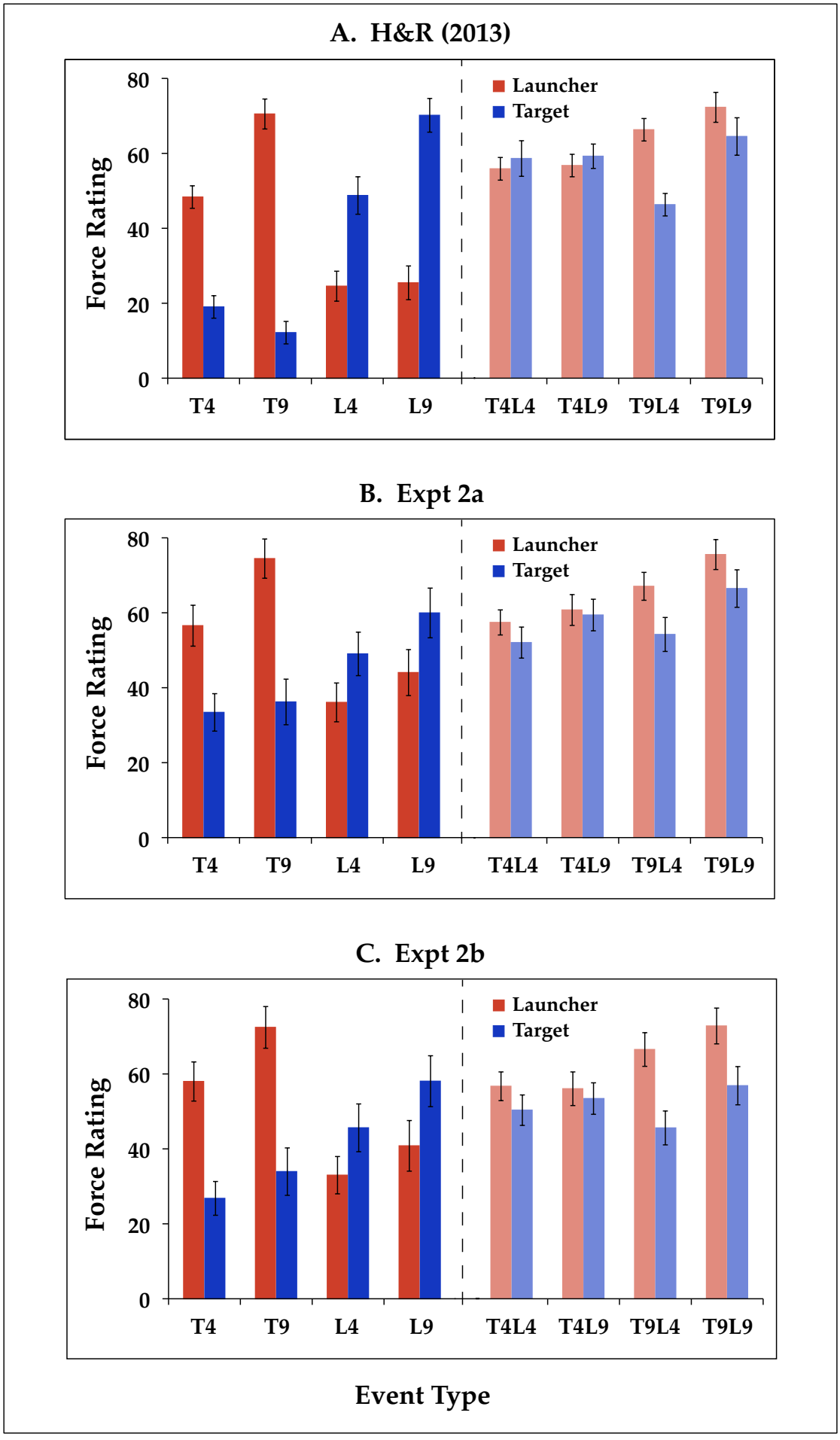

A. H&R (2013)
Force Rating
80
60
40
20
0
Launcher
Target
T4
T9
L4
L9
T4L4
T4L9
T9L4
T9L9
B. Expt 2a
Force Rating
80
60
40
20
0
Launcher
Target
T4
T9
L4
L9
T4L4
T4L9
T9L4
T9L9
C. Expt 2b
Force Rating
80
60
40
20
0
Launcher
Target
T4
T9
L4
L9
T4L4
T4L9
T9L4
T9L9
Event Type

Figure 5

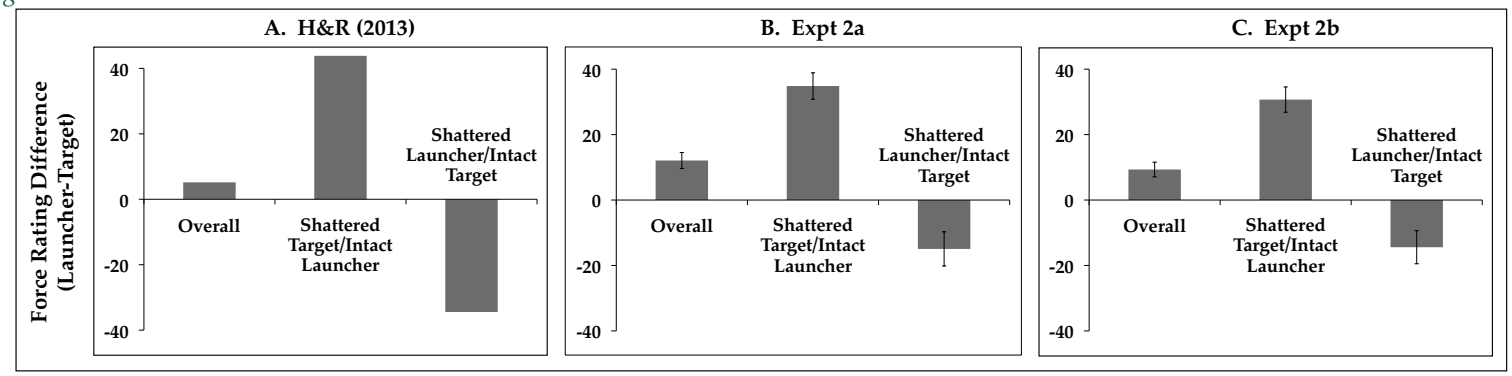
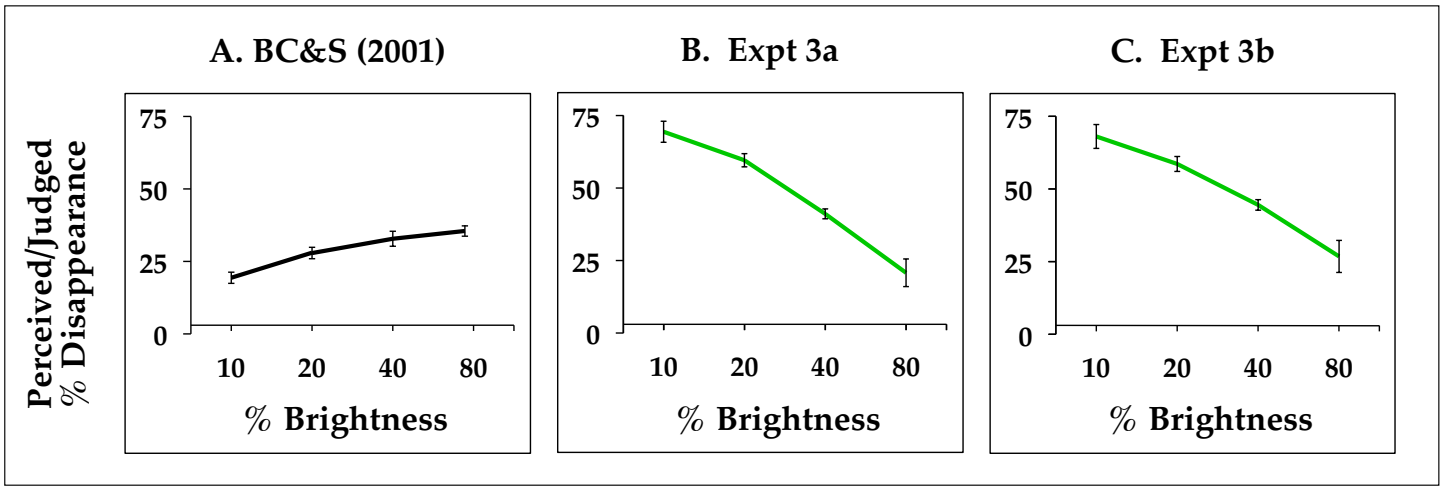

A. H&R (2013)
B. Expt 2a
C. Expt 2b
Force Rating Difference (Launcher-Target)
40
20
0
-20
-40
Overall
Shattered Target/Intact Launcher
Shattered Launcher/Intact Target

Figure 6


A. BC&S (2001)
B. Expt 3a
C. Expt 3b
Perceived/Judged
% Disappearance
75
50
25
0
10
20
40
80
% Brightness