Research Article

Psychological Science
2022, Vol. 33(2) 299–313

DOI: 10.1177/09567976211032676
www.psychologicalscience.org/PS

Scene Context Impairs Perception of Semantically Congruent Objects

Eelke Spaak, Marius V. Peelen, and Floris P. de Lange
Donders Institute for Brain, Cognition and Behaviour, Radboud University

Abstract

Visual scene context is well-known to facilitate the recognition of scene-congruent objects. Interestingly, however, according to predictive-processing accounts of brain function, scene congruency may lead to reduced (rather than enhanced) processing of congruent objects, compared with incongruent ones, because congruent objects elicit reduced prediction-error responses. We tested this counterintuitive hypothesis in two online behavioral experiments with human participants (N = 300). We found clear evidence for impaired perception of congruent objects, both in a change-detection task measuring response times and in a bias-free object-discrimination task measuring accuracy. Congruency costs were related to independent subjective congruency ratings. Finally, we show that the reported effects cannot be explained by low-level stimulus confounds, response biases, or top-down strategy. These results provide convincing evidence for perceptual congruency costs during scene viewing, in line with predictive-processing theory.

Keywords

visual perception, visual attention, prediction, perception, semantic memory, reaction time, vision, visual memory, open data, open materials

Received 9/30/20; Revision accepted 6/19/21

Objects are typically encountered in particular contexts; for example, a hair dryer is more commonly encountered in a barbershop than in a grocery store. Semantic associations between real-world scene context and objects within such scenes are well-known to facilitate perception in many circumstances: Objects are located and identified more rapidly and accurately in semantically congruent contexts than in incongruent ones (Bar, 2004; Biederman, 1972; Davenport & Potter, 2004; Kaiser et al., 2019; Oliva & Torralba, 2007). These congruency benefits are elegantly explained from the perspective of *predictive processing*, the idea that the brain is a hypothesis-testing machine (Clark, 2013; de Lange et al., 2018; Friston, 2005; Rao & Ballard, 1999): The gist of a (natural) scene induces a prior expectation over particular objects common to such a scene, and stimuli that are likely under that prior are easily integrated with it in order for the brain to arrive at a coherent representation.

Such an account explains benefits in cases in which the scene-induced expectation is relevant to the task at hand: Observers who are asked to locate a computer mouse in a scene of a desk will naturally look for it next to the keyboard and will thus more quickly find it than if it were presented in a scene of a kitchen countertop. Similarly, in a brief or degraded presentation of such scenes, an oval blob next to a keyboard-shaped blob will more readily be identified as a computer mouse than such a blob in incongruent surroundings. However, according to predictive-processing theories, it is precisely *incongruent* objects that warrant closest inspection, not congruent ones (specifically, *high-precision* incongruent objects warrant the closest inspection). The inferred identity of a congruent object is easily integrated with the prior induced by the scene gist, whereas the inferred identity of an incongruent object elicits a larger prediction error. To "resolve" this error (or, equivalently, to leverage the likely high information content afforded by the source of this error signal), observers

Corresponding Author:
Eelke Spaak, Radboud University, Donders Centre for Cognitive Neuroimaging
Email: eelke.spaak@donders.ru.nl

should associate incongruent objects with more extended processing before integration with the scene-induced prior is possible. It has been demonstrated that the amount of processing influences the level of subjective awareness (Anzulewicz et al., 2015; Windey et al., 2013). Given this assumption and the above reasoning, we hypothesized that in crowded natural scenes with a clear gist-induced prior that is not directly relevant for the current behavioral goals and with multiple objects to be explored ("sampled"), objects in congruent surroundings may be sampled less and therefore perceived less strongly, less saliently, than those in incongruent surroundings.

Researchers studying change detection have reported such context congruency costs: Observers are slower to detect changes in objects when these objects are embedded within congruent contexts, compared with incongruent ones (Hollingworth & Henderson, 2000; LaPointe et al., 2013; Mack et al., 2017). Additionally, it has been reported that, during free viewing, observers tend to fixate earlier on incongruent than on congruent objects (Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Underwood et al., 2007), and other indices of attentional allocation point in the same direction (Gordon, 2004).

However, this reported evidence for congruency costs (or, equivalently, incongruency benefits) is not clear-cut. First, several studies have failed to replicate the earlier fixation latencies for incongruent objects (De Graef et al., 1990; Henderson et al., 1999), whereas others have reported the effect only for visually non-salient objects (Underwood & Foulsham, 2006). In general, low-level visual saliency has been described as a potentially confounding factor in the research on attentional attraction by semantic incongruence (Underwood & Foulsham, 2006; Võ & Henderson, 2009). A second issue that has received less attention (though see Hollingworth & Henderson, 2000) but may be equally grave is that congruency costs might reflect a *strategic* effect: If an incongruent object is present, in many cases it will be task relevant (e.g., the changing object in change detection or something specifically memorable in a memory task), making it beneficial for participants in the experiment to search for incongruent objects in general (leading to an observed congruency cost). Finally, and perhaps most importantly, congruency costs have mainly been demonstrated through latency differences (e.g., change-detection latency) rather than through unbiased measures of perception. Response latency is influenced by multiple factors, including decisional and response biases. Previously reported congruency costs may thus reflect such biases rather than reduced perceptual encoding of congruent objects.

An intriguing possible implication of the influential idea of neural predictive processing is the existence of congruency costs in purely perceptual (i.e., nonsemantic) tasks, yet this hypothesis has not been tested directly. In the present study, we set out to perform this test while taking care of all three concerns voiced above. Importantly, we examined semantic congruency costs in a discrimination task probing object-level (i.e., exemplar) perception free from stimulus, response, semantic, and task-strategic confounds, in addition to a classic change-detection setting. In brief, across these two behavioral experiments, with a total of 300 participants, we found that congruency costs (a) are evident in change detection even with a fully balanced stimulus set, (b) generalize to a more directly perceptual identification task, (c) persist even when attending to incongruent objects is strategically disadvantageous, and (d) are explained by the subjective level of object–scene consistency.

Statement of Relevance

The theory of the "Bayesian brain," the idea that our brain is a hypothesis-testing machine, has become influential over the past decades. Particularly influential formulations are theories of predictive processing. Such theories may entail that stimuli that are expected, for instance because of the context in which they appear, generate a weaker neural response than unexpected stimuli. Scene context correctly "predicts" congruent scene elements, which should result in lower prediction error. Our study tested this important, counterintuitive, and hitherto not fully tested hypothesis. We found clear evidence in favor of it and demonstrate that these "congruency costs" are indeed evident in perception and not limited to one particular task setting or stimulus set. Because perception in the real world is never of isolated objects but always of entire scenes, these findings are important not just for the Bayesian brain hypothesis but for our understanding of real-world visual perception in general.

Method

Stimuli, task, and experimental design

Experiment 1: change detection. This experiment was a version of the classic change-detection "flicker" task (Rensink et al., 1997) and is depicted in Figure 1a. Participants were instructed to detect changes between successive displays of the same scene. Each trial started with an empty screen (500 ms), followed by a fixation button labeled "Go" in the center of the screen that participants

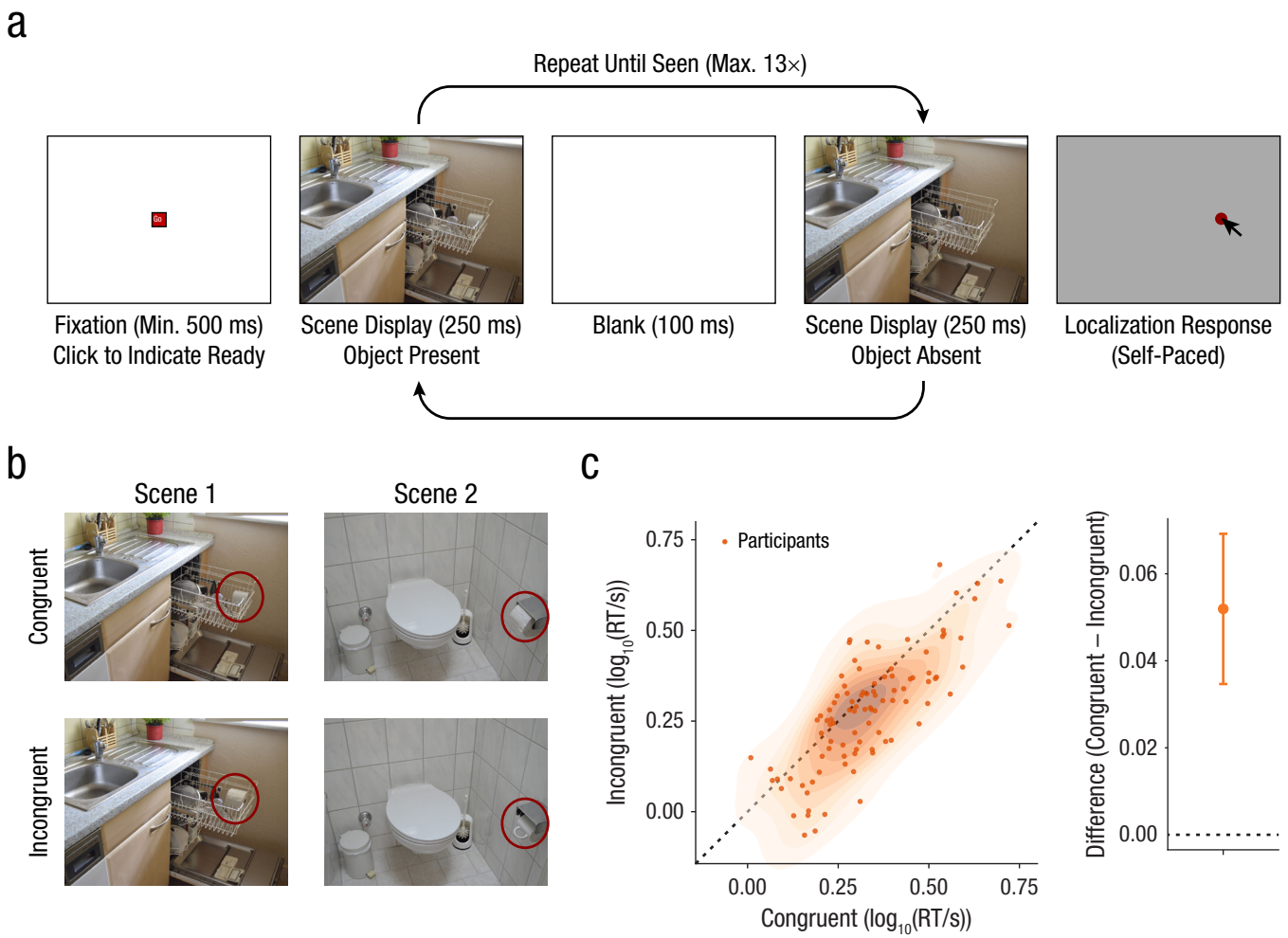

Fig. 1. Design and key results for Experiment 1. The structure and timeline of a single experimental trial is shown in (a). This was a typical "flicker" change-detection task, followed by a localization response. Participants clicked the "Go" button to initiate visual stimulation. Stimulation consisted of an object-present scene, followed by a blank screen and then an object-absent scene. This sequence was repeated a maximum of 13 times or until participants indicated that they had detected the change (in this scene, the cup disappeared from the dishwasher). After the detection response, a gray rectangle of the same dimensions as the scene stimulus appeared, and participants had to click where they had detected the changed object. Example stimuli (b) are shown for two scenes, separately for conditions in which the to-be-detected object was congruent and incongruent with the context in which it appeared (Öhlschläger & Võ, 2017). Red outlines (not shown to participants) indicate the key object (cup/toilet-paper roll in dishwasher vs. cup/toilet-paper roll in toilet-paper holder). Reaction time (RT) results are shown in (c). The scatterplot shows RT on congruent trials and RT on incongruent trials for all participants. Shaded areas indicate the density of the data (darker = denser). The graph on the right shows the mean difference between RT on congruent and incongruent trials (congruent – incongruent); the error bar shows the 95% confidence interval.

had to click to initiate visual stimulation. Requiring a mouse click in the center of the screen ensured that participants were always fixating the center at stimulus onset. Stimulation consisted of an object-present scene for 250 ms, followed by a 100-ms blank screen and then an object-absent scene. This sequence was repeated a maximum of 13 times or until participants indicated that they had detected the change by pressing the space bar. After the detection response, a gray rectangle of the same dimensions as the scene stimulus appeared, and participants had to click where they had detected the changed object. This ensured task compliance and prevented blind, rapid pressing of the space bar.

Stimuli were taken from a recently published, fully balanced stimulus database called SCEGRAM (Öhlschläger & Võ, 2017). We used all 62 scenes from the conditions labeled "CON" and "SEM" in the SCEGRAM database (we refer to these conditions here as the "congruent" and "incongruent" conditions, respectively), as well as the corresponding object-absent scenes. This database comprises pairwise balanced stimuli matched in lower level visual features. Each key object occurred in both a congruent context (e.g., a cup in a dishwasher) and an incongruent context (e.g., a cup in a toilet-paper holder) and was matched with another key object similar in shape and orientation that had complementary

congruency mapping (e.g., a toilet-paper roll in either a toilet-paper holder [congruent] or a dishwasher [incongruent]; see Fig. 1b).

Each participant completed 62 trials. Half of these contained congruent scenes, and half contained incongruent scenes. This mapping was counterbalanced across participants. The main dependent variable was change-detection reaction time (RT); localization error was a secondary dependent variable. Localization error was defined in percentage of the target object (e.g., a click located 1 cm away from the center of a 1 cm × 1 cm object would be assigned an error of 100%), and any errors of less than 50% indicated that the observer clicked within the bounds of the target object. Specifically, localization error was defined as follows:

$$\mathrm{err}_x = \frac{x_{\mathrm{click}} - x_{\mathrm{object\ center}}}{w_{\mathrm{object}}} \times 100\%$$

$$\mathrm{err}_y = \frac{y_{\mathrm{click}} - y_{\mathrm{object\ center}}}{h_{\mathrm{object}}} \times 100\%$$

$$\mathrm{loc\ err} = \sqrt{\mathrm{err}_x^2 + \mathrm{err}_y^2}$$

For all analyses of RT, we focused only on those change-detection responses for which the subsequent localization error was 100% or less.

Experiment 2: object identification. For this experiment, instead of having to detect a change, participants were instructed to attentively look at each scene and then make an identification judgment about an object in that scene. Each trial started with a fixation cross (800–1,000 ms, randomly drawn from a uniform distribution) followed by a scene display (2.5 s). After another fixation cross (500 ms), participants were given a two-alternative forced-choice (2AFC) response prompt, which lasted a maximum of 2.5 s or until one of the response keys was pressed (see Fig. 2a). The 2AFC prompt always consisted of the target item that was present in the scene (again taken from the SCEGRAM stimulus database) as well as a lure item selected through an Internet search. The lure was always from the same category as the target and similar in shape but a clearly different exemplar.

As in Experiment 1, scenes could occur in a congruent or an incongruent condition. Whether a scene was congruent or incongruent was determined by the nature of the key object in the scene (e.g., the cup or toilet-paper roll in Figs. 1b and 2). For Experiment 2, we added a probe (key vs. other) factor, which governed whether, on a given trial, the probe was about this key object or about another object in the same scene. The probed object in probe-other trials was always congruent with the surrounding context (see Fig. 2b). The rationale for including probe-other trials is twofold. First, this allowed us to control for the strategic concern mentioned in the introduction; that is, if an incongruent object was present, this was no longer necessarily the relevant object (Hollingworth & Henderson, 2000). Second, the presence of an irrelevant congruent/incongruent object might draw attention away from the probed target object; this effect should be detectable.

From the participant's perspective, there was no subjective difference between a congruent/probe-key trial and a congruent/probe-other trial: On both trial types, there were only congruent items present in the scene, and the participant was later probed about one of them. This distinction nevertheless is important because of the design of the stimulus set. Any given congruent/probe-key stimulus was matched with an incongruent/probe-key stimulus (i.e., the same scene with either a congruent or an incongruent item, where these key items were matched for location, shape, and size). Similarly, any given congruent/probe-other stimulus was matched with an incongruent/probe-other stimulus (i.e., again the same scene with a matched congruent/incongruent item present but where the probe was now about another, nonmanipulated object that was identical across the two levels of congruency). This crossed Probe × Congruency manipulation is very similar to the design used in previous work by Hollingworth and Henderson (2000).

Stimuli were again counterbalanced across participants, each of whom again completed 62 trials. Trials were again 50% congruent and 50% incongruent. Probe × Congruency together formed a 2 × 2 factorial design, but trial counts per cell were deliberately not fully equalized for all participants. Instead, we introduced a between-subjects factor, $p(\text{probe key} = 1 \mid \text{incongruent} = 1)$, or $p(\text{probe key} \mid \text{incongruent})$ for short, which governed the percentage of incongruent trials that were probe-key trials—that is, given that an incongruent object was present in the scene, $p(\text{probe key} \mid \text{incongruent})$ determined the probability that this object was the task-relevant one. This factor took on values of 17%, 33%, 50%, 67%, and 83%. For values of the between-subjects factor other than 50% (the fully across-subjects counterbalanced case), a randomly chosen subset (per participant) of incongruent trials was switched from probe-key to probe-other trials or vice versa. This factor allowed us to quantify the degree to which behavioral costs/benefits were a consequence of task strategy.

The main dependent variable for this experiment was 2AFC accuracy; RT was of secondary interest. For RT analyses, we focused only on trials with correct responses.

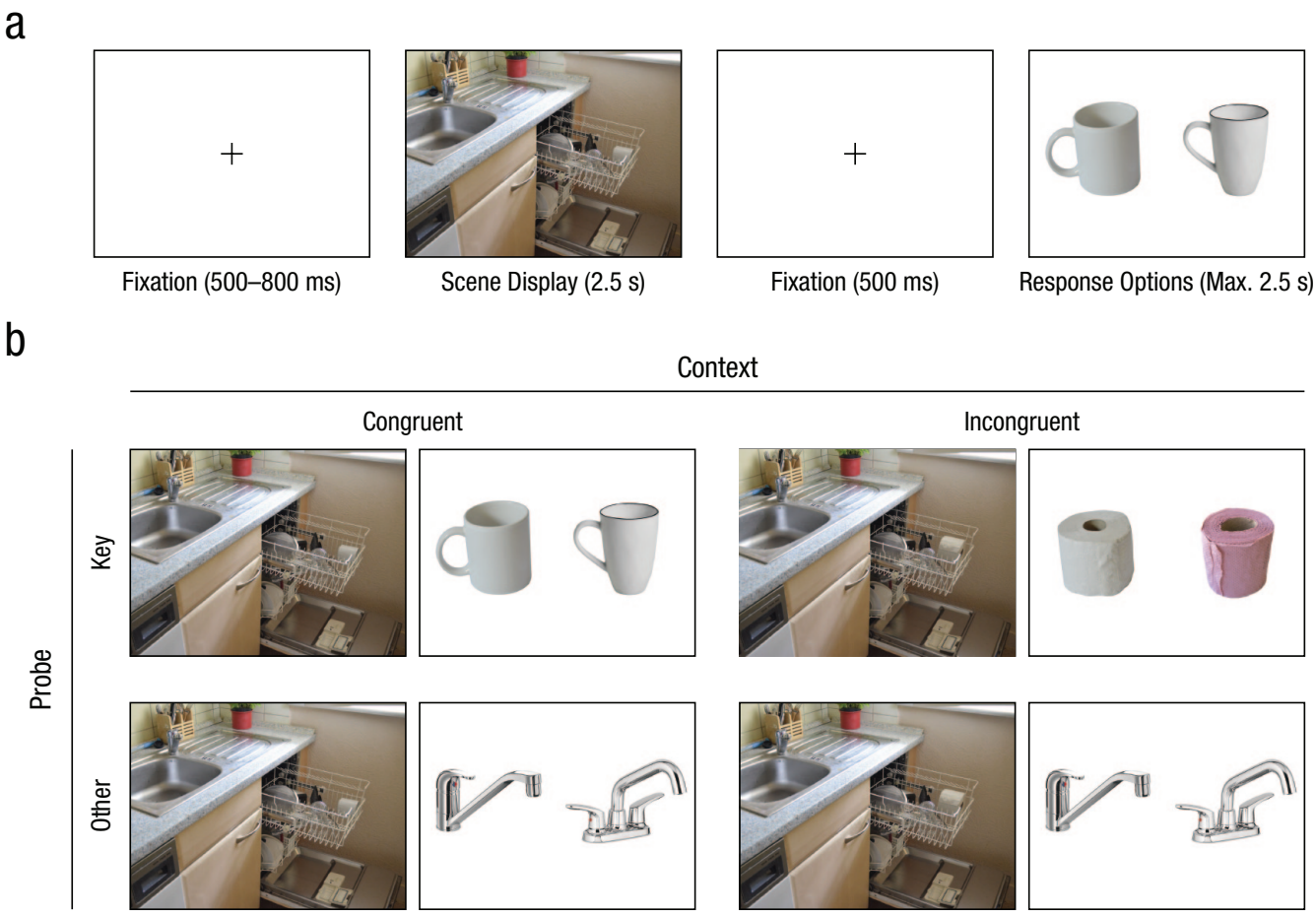

Fig. 2. Design for Experiment 2. The structure and timeline of a single experimental trial is shown in (a). Participants were shown a natural (indoor) scene and afterward were asked to make a two-alternative forced choice to identify which object had been present in the scene. The two within-subjects manipulations are illustrated in (b). Scenes could be either congruent (left) or incongruent (right), and the probed objects could be either the key object (top) or another object (bottom). Whether a scene was congruent or incongruent was determined by the nature of the key object in the scene (e.g., a toilet-paper roll was congruent with a bathroom scene but incongruent with a kitchen scene).

Participants, data inclusion, and statistical power

The experiments were conducted online using the Gorilla platform (Anwyl-Irvine et al., 2020), and participants were recruited through the Prolific platform (https://www.prolific.co/). The study was approved by the local ethics committee (Commissie Mensgebonden Onderzoek Arnhem-Nijmegen, Radboud University Medical Center) under the general ethical approval for online studies for the Donders Centre for Cognitive Neuroimaging.

To increase the signal-to-noise ratio of our data set, we made an a priori decision to remove outliers, which might be especially expected in an online setting. We defined outliers as participants who scored greater than 2.5 standard deviations away from the mean for either dependent variable. Outlier detection was performed on overall scores regardless of condition. For Experiment 1, we recruited 100 participants (49 female, 51 male; age: M = 27.35 years, SD = 5.81), of which three were classified as outliers and removed (see Fig. S1 in the Supplemental Material available online). The recruitment target was chosen for convenience.

Effect sizes for similar experiments unfortunately have not been reliably reported in the literature, and where they are reported, they vary widely. Three studies comparing change-detection RTs between congruent and incongruent contexts allowed us to compute effect-size d in two ways. The first is from reported paired-samples t statistic and sample size ($d = t/\sqrt{n}$): d = 1.6 (LaPointe et al., 2013) and d = 0.55 (Mack et al., 2017). The second is from reported mean square error and sample size (Thalheimer & Cook, 2002): d = 1.1 (Hollingworth & Henderson, 2000). However, for reasons noted in the introduction, these effect sizes might be overestimating a potential true effect, and we therefore

cannot assume these to hold for our effect of interest per se. For this reason, we calculated power for an effect that we minimally wanted to be sensitive to. Note that, as mentioned, the sample size for Experiment 1 was chosen for convenience, and we provide the power calculation here only for information. The sample of 97 participants yielded a power of 99.8% for a paired contrast (two tailed) in the case of a medium effect size (d = 0.5) or 49.6% power in the case of a weak effect size (d = 0.2), based on an a priori Type I error rate (α) of .05.

For Experiment 2, we recruited 200 participants (94 female, 106 male; age: M = 29.85 years, SD = 6.16), resulting in 40 participants for each level of the between-subjects factor. This sample size was chosen to ensure at least 80% power (two tailed) to detect an effect of medium size ($d \geq 0.5$) within each level given 34 participants per level, plus margin for potential rejection of unreliable data. Six participants were classified as outliers and removed (see Fig. S6 in the Supplemental Material). In addition to the a priori power considerations regarding the paired comparisons, the power to detect the effect of the between-subjects factor is relevant here (which we provide only for information and did not use when determining the sample size). The resulting sample of 194 participants yielded a power of 87.8% (one tailed) to detect a weak correlation (r = .2). None of the participants in Experiment 2 had participated in Experiment 1. All power calculations were performed using G*Power (Version 3.1; Faul et al., 2009).

In addition to detecting outlying participants, we also screened for outlying experimental stimuli (i.e., scenes), again in a condition-agnostic fashion. There were two outlying items in Experiment 1 and zero outlying items in Experiment 2. These items were discarded from further analysis.

Data analysis

All analyses were performed using custom-written scripts in *Python* (Van Rossum & Drake, 1995) using the *NumPy* (van der Walt et al., 2011), *SciPy* (Virtanen et al., 2020), *Pingouin* (Vallat, 2018), *Pandas* (McKinney, 2010), *Matplotlib* (Hunter, 2007), *Seaborn* (Waskom et al., 2020), *PyMC3* (Salvatier et al., 2016), *ArviZ* (Kumar et al., 2019), and *Bambi* (Yarkoni & Westfall, 2016) libraries. For software package versions, refer to the GitHub link in the Open Practices statement at the end of this article.

For all pairwise comparisons and correlations, in addition to frequentist statistics such as t values, we report Bayes factors (BFs) quantifying how much more likely the data are under the alternative hypothesis than under the null hypothesis (BF_{10}). BFs were estimated analytically using the following priors for t tests: a Cauchy prior on effect size and a Jeffreys prior on variance, resulting in a Jeffreys-Zellner-Siow (JZS) BF (Jeffreys, 1998; Rouder et al., 2009; Zellner & Siow, 1980). The scale parameter for the Cauchy prior on effect size (r) was set to .33, corresponding to an 80% a priori probability that the observed effect size (d) lies between −1 and +1 (or, equivalently, between 0 and ±1 for a directional test; Schmalz, 2019). We chose this on the basis of published effect sizes in similar studies (see above). For the three primary paired contrasts of congruent versus incongruent, we explored the resulting t-based BF_{10} values across different plausible levels of Cauchy scale values to verify that our conclusions did not critically depend on the choice for the default. This control analysis is shown in Figure S12 in the Supplemental Material. BF_{10}s for correlations were calculated according to the scheme outlined by Ly et al. (2016), with noninformative default priors (i.e., $\kappa = 1$) corresponding to a uniform prior distribution over the interval [−1, +1]. Although specifically formulated for Pearson correlations, the same BF_{10} calculation was used for Spearman correlations as well, because Spearman correlation is equivalent to Pearson correlation on rank-transformed data (Myers & Well, 2003, p. 508).

In addition to the simple paired comparisons and correlations, we report results from Bayesian hierarchical generalized linear models (also known as mixed-effects or multilevel models) with full random-effects structure. For details on the models and sampling scheme, see Note 1 in the Supplemental Material. Results from these analyses are primarily summarized using 94% highest-density intervals (HDIs).

Before conducting all statistical analyses (including outlier rejection), we $\log_{10}$-transformed bounded variables (RT, localization error) to improve normality and stabilize variance. For all tests with a priori directional hypotheses, we report one-tailed p and BF_{10} values; tests without a priori directional hypotheses were conducted using two-tailed values.

Results

Congruency costs in change detection with controlled stimuli

A sample of 100 volunteers participated online in Experiment 1, which was a version of the classic "flicker" change-detection paradigm (Rensink et al., 1997; see Fig. 1a) with an added localization response. Importantly, scene changes could occur with either a congruent or an incongruent object, where low-level similarities between conditions were matched as much as possible (Öhlschläger & Võ, 2017; see Fig. 1b) and stimuli were counterbalanced across participants.

Change-detection RTs were well within the maximum of 7.8 s (M = 1,576 ms, SD = 234; see Fig. S1), indicating that participants were able to perform the task successfully. This was further corroborated by the localization-error scores, which demonstrate that participants on average were able to locate the item correctly (M = 26.8%, SD = 9.7, where a value < 50% indicates a click inside the key object; see Fig. S1). Only 2.10% of total responses had a localization error greater than 100% and were excluded from all RT analyses.

The key hypothesis that this experiment was designed to test was that change-detection performance suffers when objects are presented in congruent contexts, compared with when objects are presented in incongruent contexts. Specifically, as operationalized here, RTs for congruent trials should be longer than for incongruent trials. This is exactly what we found, $t(96) = 4.98$, $p < .001$, $d = 0.51$, $BF_{10} = 10{,}895.21$; difference: $M = 0.052$ $\log_{10}$(RT/s), 95% confidence interval (CI) = [0.035, 0.070]; congruent: M = 1.63 s, incongruent: M = 1.51 s, difference = 0.12 s (see Fig. 1c). Additionally, we found that localization errors were larger on congruent than on incongruent trials, $t(96) = 2.44$, $p = .008$, $d = 0.25$; difference $M = 0.020$ $\log_{10}$, 95% CI = [0.0064, 0.033] (see Fig. S2 in the Supplemental Material), though the evidence for this effect was only moderate ($BF_{10} = 5.93$).

Although the stimulus set we used was highly controlled, stimuli were still instances of natural scenes. This improved the ecological validity of the experiment, but it also meant that stimuli were necessarily a sample of all possible scenes that could have been used. This led to variation in the effect across experimental items (see Fig. S3 in the Supplemental Material). In addition to the null-hypothesis test described above, we therefore conducted a fully Bayesian hierarchical analysis to account for this. In this case, the Bayesian test was analogous to a paired contrast with random intercepts and slopes across both participants and stimulus items. The results of this analysis corroborated the conclusions of the null-hypothesis test: RTs were slower for congruent than for incongruent trials, coefficient posterior $M = -0.058$ $\log_{10}$(RT/s), 94% HDI = [−0.10, −0.010] (see Fig. S4 in the Supplemental Material). Furthermore, the Bayesian analysis allowed us to generalize our conclusion further: Across both the population from which our participants were drawn and the population from which our stimuli were drawn, we can be 98.71% certain that the RT effect was nonzero. The Bayesian analysis of localization error also yielded a corroboration of the analogous null-hypothesis test, albeit a weaker one ($M = -0.021$ $\log_{10}$, 94% HDI = [−0.046, 0.0061]; see Fig. S5 in the Supplemental Material), with 92.77% probability that the effect was lower than zero.

In summary, Experiment 1 provided strong evidence that change detection was impaired when the changing object appeared within a semantically congruent, rather than semantically incongruent, context. We extend the existing literature by showing that this effect persisted even for controlled, matched stimuli.

Congruency costs extend to perceptual identification

A key motivation for the present research was to investigate whether congruent surroundings have consequences for the perception of the congruent or incongruent key object itself. We therefore conducted a second experiment using the same stimuli as in Experiment 1 but with a different, more directly perceptual task. Two hundred volunteers participated in Experiment 2, in which they had to identify one of two similar objects as having been present in a previously presented scene (see Fig. 2a). Context was again included as a factor (congruent vs. incongruent), and we additionally included the factor probe (key vs. other), resulting in the addition of trials in which an incongruent object was present but was not the target item (see Fig. 2b).

Overall, 2AFC accuracy was clearly above chance level (M = 61.44%, SD = 7.72), and RTs were well within the maximum of 2.5 s (M = 1,226 ms, SD = 216; see Fig. S6), indicating that participants were able to perform the task successfully.

We found that 2AFC accuracy was significantly lower for congruent trials than for incongruent ones when we focused on the probe-key condition, $t(193) = -4.49$, $p < .001$, $d = 0.32$, $BF_{10} = 2{,}733.00$; difference: M = −6.18%, 95% CI = [−8.46, −3.90] (see Fig. 3). There was no difference between congruent and incongruent trials in the probe-other condition, $t(193) = -0.23$, $p = .591$, $d = 0.02$; difference: M = −0.30%, 95% CI = [−2.45, 1.85] (see Fig. 3), and the data were about 3 times more likely under the null hypothesis of no difference ($BF_{10} = 0.34$). RT data showed a very similar pattern: Responses were slower for congruent than for incongruent trials within the probe-key condition, $t(193) = 4.21$, $p < .001$, d = 0.30, $BF_{10} = 677.40$; difference $M = 0.053$ $\log_{10}$(RT/s), 95% CI = [0.032, 0.074]; congruent: M = 1.15 s, incongruent: M = 1.09 s, difference: M = 0.060 s (see Fig. S7 in the Supplemental Material), but there was no such difference for the probe-other condition, $t(193) = -1.03$, $p = .152$, $d = 0.07$; difference: $M = -0.013$ $\log_{10}$(RT/s), 95% CI = [−0.035, 0.0080]; congruent: M = 1.21 s, incongruent: M = 1.24 s, difference M = −0.032 s (see Fig. S7), although the evidence for a null effect here was anecdotal at best ($BF_{10} = 0.54$).

Because Experiment 2 had a 2 × 2 design, it was important to formally test the interaction between the two factors. Additionally, as in Experiment 1, we sought to test the generalization of the observed effects to the

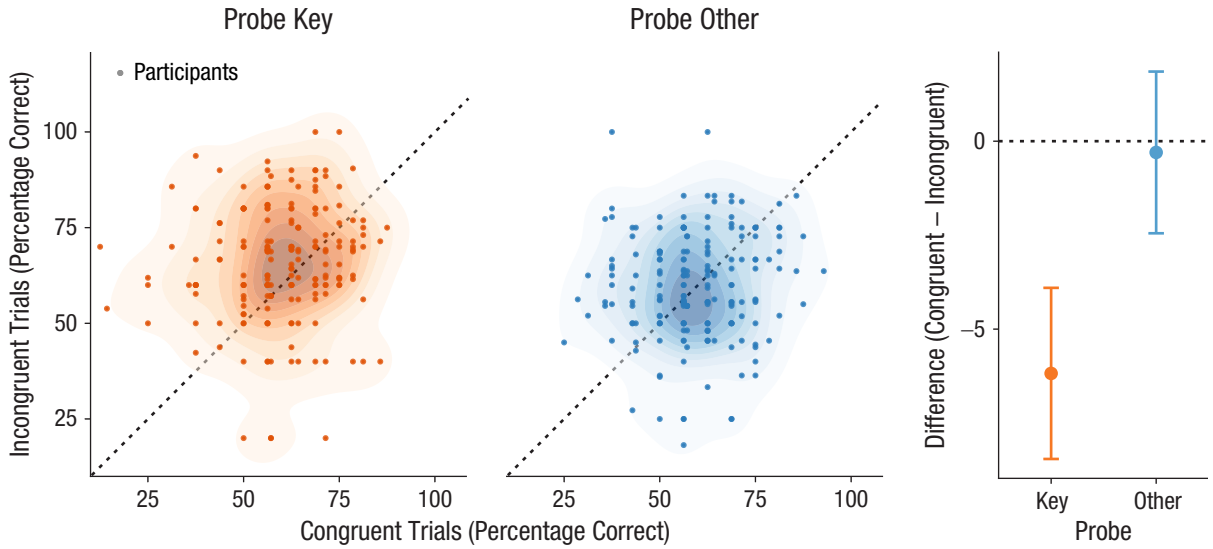

Fig. 3. Two-alternative forced-choice accuracy for Experiment 2. The scatterplots show accuracy in congruent trials and incongruent trials, separately for the probe-key and probe-other conditions. Shaded areas indicate the density of the data (darker = denser). The graph on the right shows the mean difference in accuracy between congruent and incongruent trials for each probe condition; error bars show 95% confidence intervals.

population not just of participants but of experimental items as well (for effect spread over items, see Fig. S8 in the Supplemental Material). To accomplish these goals, we again conducted a hierarchical Bayesian (logistic) regression analysis of 2AFC accuracy with the key experimental effect captured by the interaction parameter. We found clear evidence for an interaction effect (M = 0.39, 94% HDI = [0.12, 0.65]; see Fig. S9 in the Supplemental Material) and a 99.66% probability that the parameter exceeded zero, indicating that participants indeed were more accurate on incongruent trials specifically when probed about the (incongruent) key object and not when probed about another.

The RT data showed a very similar pattern: Participants were faster on incongruent trials, specifically in the probe-key condition (interaction parameter: M = −0.064 $\log_{10}$(RT/s), 94% HDI = [−0.11, −0.014]; see Fig. S10 in the Supplemental Material), with a probability of 99.24%. RT analysis additionally revealed a main effect of probe, M = −0.060 $\log_{10}$(RT/s), 94% HDI = [−0.097, −0.024], indicating that participants were faster to respond on probe-key than on probe-other trials (probability 99.88%). It is possible that the stimuli in the probe-other condition were more difficult than those in the probe-key condition, but given the convincing absence of a main probe effect in 2AFC accuracy (94% HDI = [−0.19, 0.30]), we cannot conclude this with certainty.

Taken together, we can conclude with considerable confidence that congruency costs are not limited to perceptually indirect measures such as change detection or spatial attention allocation, but they extend to object-exemplar identification and thus have genuine perceptual consequences (for an additional result regarding the possible attentional locus of this effect, see Note 2 in the Supplemental Material).

Congruency costs are not explained by task strategy

In Experiment 1, as in the majority of previous research on object congruence, if an incongruent item was present in a scene, it was always the task-relevant item. Any congruency costs (or incongruency benefits) might therefore be explained by participants adopting the strategy of always searching for an incongruent object and paying full attention to it. This strategy is effective only on incongruent trials and not on congruent ones, hence potentially causing a general congruency cost. In Experiment 2, we included the probe factor specifically to ensure that incongruent items, when present, were not automatically task relevant (see Fig. 2b). This factor by itself, in a 2 × 2 balanced design, already ensured that p(probe key = 1 | incongruent = 1) was reduced to 50% (from the typical 100%). For the above analysis, p(probe key | incongruent) indeed was 50% on the aggregate, yet we observed that congruency costs persist. Therefore, on the basis of the above, we can already conclude that congruency costs are not exclusively observed in settings in which p(probe key | incongruent) was equal to 100%.

It is possible that congruency costs, although not abolished, were still attenuated for lower values of

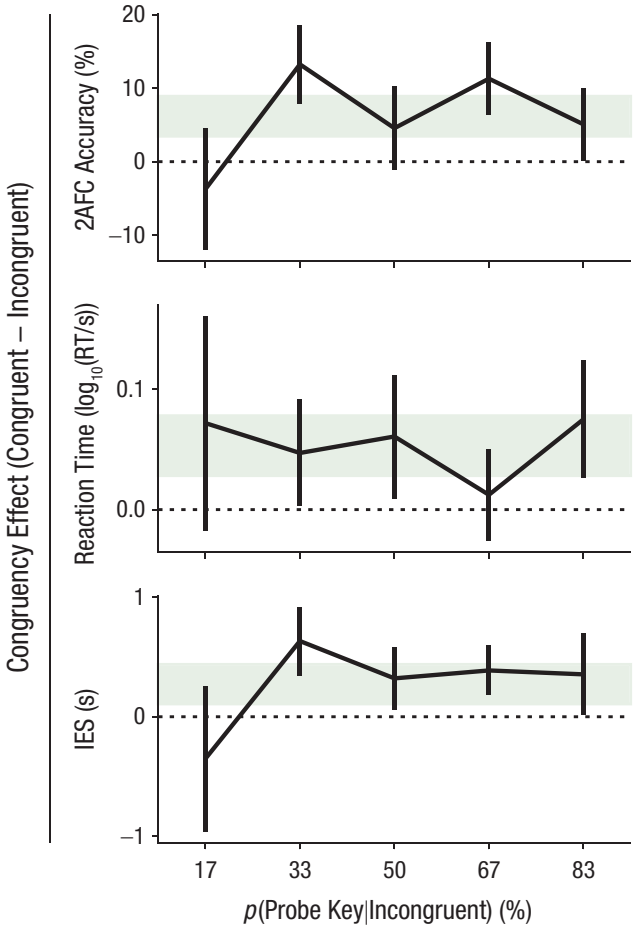

Fig. 4. Congruency effects in Experiment 2. The mean congruency difference score (congruent – incongruent) is shown for each level of *p*(probe key | incongruent): the two main dependent variables (two-alternative forced-choice [2AFC] accuracy and reaction time [RT]), as well as a combined metric, inverse-efficiency score (IES). Error bars show 95% confidence intervals, and shading reflects the 95% confidence interval of the effect across the entire sample.

p(probe key | incongruent). If this were the case, then the strategic concern described above might still be an issue. For Experiment 2, we manipulated *p*(probe key | incongruent) across participants to test to what extent the potential confound of task strategy might explain observed congruency costs. We found no effect of *p*(probe key | incongruent) on task effects in either 2AFC accuracy, Spearman's $r(192) = .074$, $p = .30$, 95% CI = [−.067, .21] (see Fig. 4), or RT, $r(192) = -.023$, $p = .75$, 95% CI = [−.16, .12], and the data were about 7 times more likely under the null hypothesis of no correlation for accuracy ($BF_{10} = 0.15$) and about 11 times more likely under the null hypothesis for RT ($BF_{10} = 0.094$). Visual inspection of the data (see Fig. 4) showed that fluctuations in accuracy across *p*(probe key | incongruent) were accompanied by opposite fluctuations in RT, perhaps suggesting variations in speed/accuracy trade-off. To account for this, we additionally computed the inverse-efficiency score (Vandierendonck, 2017), which also did not show an effect of *p*(probe key | incongruent), $r(192) = .045$, $p = .53$, 95% CI = [−.096, .19], and the data were about 9 times more likely under the null hypothesis ($BF_{10} = 0.11$).

We finally note that the cell *p*(probe key | incongruent) = 17% contained relatively few incongruent/probe-key trials per participant; thus, the effect of congruency was estimated less reliably for these participants (e.g., this is evident in the increased error bars in Fig. 4 for this level of the independent variable). For both accuracy and inverse-efficiency score, it may appear as though there was no congruency effect for the 17% cell (or even a negative one), but the data are inconclusive on this point ($BF_{10} = 0.91$ for accuracy; $BF_{10} = 0.90$ for inverse-efficiency score).

In summary, the persistence of congruency costs on probe-other trials, and in particular the absence of a modulation of congruency costs in response to increasing the task relevance of incongruent objects, is clear evidence that the congruency-cost phenomenon cannot be explained by task-strategic considerations.

Change detection and exemplar identification tap into overlapping effects

A natural question to ask is whether the congruency-cost effects identified in both experiments are related. The two experiments were performed using two different samples of participants, yet they employed the same stimulus set. Therefore, to answer this question, we examined this relationship across experimental items. We found that, indeed, the experimental effects were related between the two experiments. For the main dependent variables of change-detection RT (Experiment 1) and 2AFC identification accuracy (Experiment 2), we found clear evidence for a negative correlation across items for congruency difference scores, $r(58) = -.35$, $p = .003$, 95% CI = [−.55, −.10], $BF_{10} = 11.46$ (see Fig. 5). (Note that the negative direction of the effect is explained by positive [or positive changes in] accuracy values indicating better performance, whereas positive [or positive changes in] RT values indicate worse performance.) Correlations between the other dependent variables corroborated this finding: Change-detection-localization performance was highly correlated with identification RT, $r(58) = .40$, $p < .001$, 95% CI = [.16, .59], $BF_{10} = 43.24$, whereas the other pairwise correlations were not significant and presented evidence ranging from anecdotal to inconclusive: Experiment 1 RT × Experiment 2 RT: $r(58) = .10$, $p = .21$, 95%

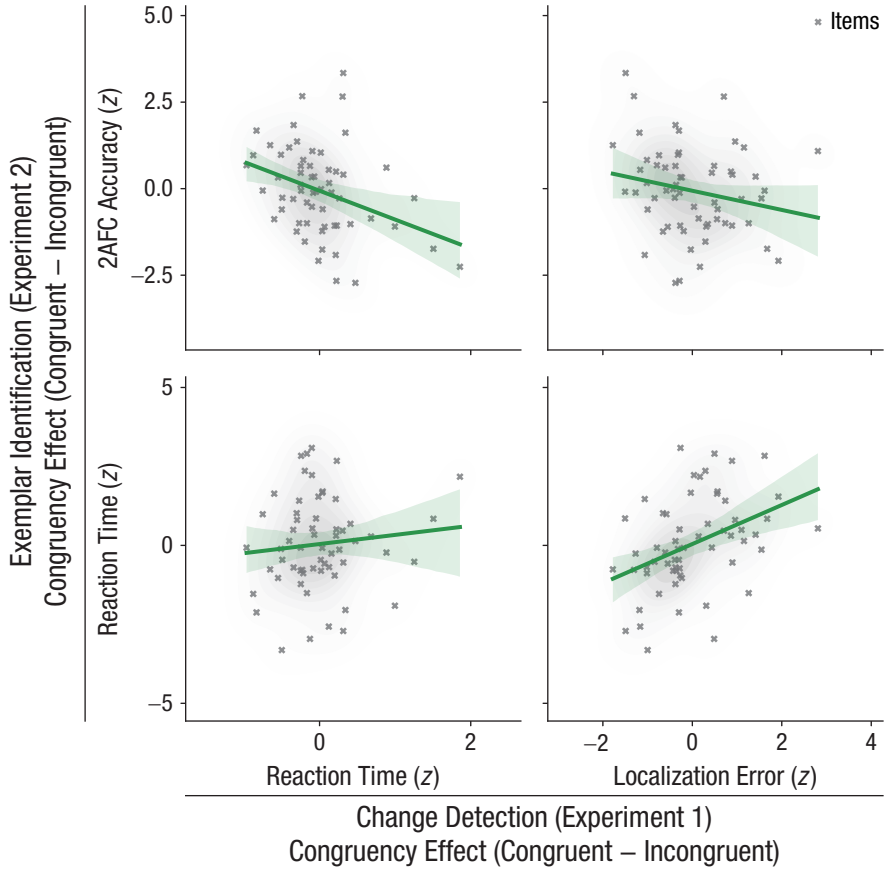

Fig. 5. Relation between the congruency effects for exemplar identification (Experiment 2) and change detection (Experiment 1). Two-alternative forced-choice (2AFC) accuracy (top row) and reaction time (bottom row) in Experiment 2 are shown as a function of reaction time (left column) and localization error (right column) in Experiment 1. Lines indicate best-fitting regressions of the 2AFC effect onto the change-detection effect, and the error band around each regression line indicates the 95% confidence interval. Shaded areas around the data points indicate the density of the data (darker = denser).

CI = [−.15, .35], BF_{10} = 0.34; Experiment 1 Localization Error × Experiment 2 2AFC Accuracy: $r(58) = -.21$, $p = .056$, 95% CI = [−.44, .05], BF_{10} = 1.05 (see Fig. 5).

We can thus conclude that the two perceptual tasks, change detection in Experiment 1 and exemplar identification in Experiment 2, tap into overlapping effects.

Subjective congruency ratings partly explain behavioral performance

Within the incongruent scenes, there was variation in the extent to which an object might be considered incongruent. This variation was previously established by the authors of the original publication about the stimulus set we used (Öhlschläger & Võ, 2017): A separate sample of observers independently rated each of the scene stimuli. In a final, exploratory analysis, we asked whether this variation in incongruency ratings might explain part of the congruency effects we observed. Specifically, we looked at the correlation between subjective inconsistency ratings (where higher means more inconsistent) and behavioral performance across the incongruent scenes. Here, as in the section above, we leveraged variation across experimental items rather than across participants.

Change-detection RTs in Experiment 1 were not correlated with inconsistency ratings, $r(58) = .05$, $p = .358$, 95% CI = [−.21, .30], BF_{10} = 0.12 (see Fig. 6). However, 2AFC accuracy scores in Experiment 2 were significantly correlated with inconsistency ratings, and higher subjective inconsistency corresponded to better performance, $r(60) = .32$, $p = .005$, 95% CI = [.08, .53], BF_{10} = 7.67 (see Fig. 6). Neither of the secondary dependent variables in the two experiments was significantly correlated with inconsistency ratings (both $|r|$s < .15, ps > .10, BF_{10}s < 0.2; see Fig. S11 in the Supplemental Material).

We can conclude that objects within scenes that are rated as subjectively more incongruent by independent

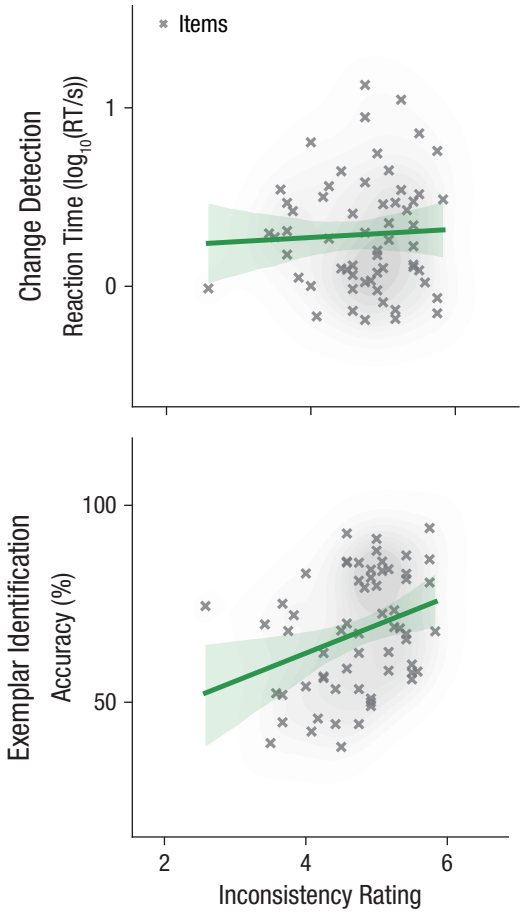

Fig. 6. Behavioral performance for Experiment 1 (change detection) and Experiment 2 (exemplar identification) as a function of subjective inconsistency rating. Subjective inconsistency ratings for each experimental item were provided by an independent sample of observers and collected by Öhlschläger and Võ (2017). Higher ratings indicate stronger inconsistency (i.e., lower object/scene consistency). Data points are for individual items (incongruent scenes only). Lines indicate best-fitting regressions, and the error band around each regression line indicates the 95% confidence interval. Shaded areas around the data points indicate the density of the data (darker = denser). RT = reaction time.

observers are easier to identify (Experiment 2), but changes in those objects are not detected faster (Experiment 1).

Discussion

In two large-sample behavioral experiments, we tested a counterintuitive consequence of predictive-processing theories of brain function: that scene-induced priors can lead to impaired perception of scene-congruent objects. In Experiment 1, we replicated the established congruency-cost effect from the change-detection literature while crucially controlling for potential stimulus confounds previously uncontrolled for. More importantly, in Experiment 2, using a task requiring object-exemplar identification, we demonstrated that congruent surroundings impair perception of key objects themselves. By manipulating the percentage of incongruent objects that were relevant, we were able to show that congruency costs are not due to participant strategy but, rather, reflect an automatic perceptual phenomenon.

Previous studies have reported slower change-detection responses for changes in congruent objects than for changes in incongruent ones (Hollingworth & Henderson, 2000; LaPointe et al., 2013; Mack et al., 2017). Eye-tracking studies have additionally reported that scene-incongruent objects are fixated earlier than congruent ones (Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Underwood et al., 2007). It has been debated to what extent these effects are truly due to *semantic* congruence or might instead be better explained by low-level visual features (e.g., local contrast) differing between conditions (Underwood & Foulsham, 2006; Võ & Henderson, 2009). In the present study, we used a stimulus database designed to be highly balanced between the semantically congruent and semantically incongruent conditions (Öhlschläger & Võ, 2017). A further advantage of these stimuli is that they were all actual photographs and did not rely on digital image-editing techniques to transplant objects from one context to another. Such digital editing, even when used carefully, might introduce local inconsistencies (e.g., in lighting) that are not exclusively semantic. The results for Experiment 1 demonstrate that congruency costs are evident in change detection even for such controlled stimuli, ruling out worries that this semantic effect might not truly be semantic at all.

An additional potential concern is that of task strategy. In a typical experimental design with congruent and incongruent trials, if an incongruent object is present in a scene, then it is very likely task relevant (e.g., the locus of change in change detection or specifically memorable in a memory task). Participants might therefore decide to always look for an incongruent object because this will be beneficial in half the trials (and thus in general). This strategy is effective only on incongruent trials, thereby leading to a behavioral benefit in that condition. We quantified this potential strategic effect in Experiment 2 by including trials on which an incongruent object was present but not task relevant. Importantly, we manipulated the percentage of trials on which a presented incongruent object was relevant across participants and found no modulation of congruency costs by this percentage. We can therefore conclude that these congruency costs arise automatically, independent from (deliberate or unconscious) strategic choices. This conclusion is in line with Hollingworth and Henderson's (2000) work, which included a similar

manipulation to our probe (key vs. other) factor. We now crucially extended this work with the manipulation of relevance probability.

As described, previous work has identified congruency costs (or, equivalently, incongruency benefits) in perceptually indirect, attentional measures. Perhaps the most important advance made by Experiment 2 over earlier work is that it probed the consequences of scene congruency for object perception itself, through a different task from that previously used, namely, exemplar identification. We used accuracy in a bias-free discrimination task as a dependent variable, thereby enabling us to test whether scene congruency affects perceptual encoding, independently of decisional and response biases. Just as in Experiment 1, we found clear evidence for congruency costs, which were furthermore related to the effect observed during change detection. Congruency costs thus not only reflect nonspecific biases or some peculiarity of change detection but also appear generally during the process of perceiving objects within scenes.

It is possible that the effects we observed in Experiment 2 were, at least partly, mediated by participants directing their attention (overtly or covertly) to incongruent objects more frequently than to congruent ones. This would be in line with some interpretations of predictive-processing theory, which posit that prediction errors elicited by incongruent objects increase the salience of these objects and thus attract attention (den Ouden et al., 2012; Feldman & Friston, 2010; see also the next paragraphs). Because we did not measure eye movements, our results do not speak directly for or against this interpretation. The absence of a congruency effect when the key object was irrelevant (probe-other trials; see also Note 2 in the Supplemental Material) provides some circumstantial evidence against spatial attention being the only factor at play here, but further research is needed to determine whether congruency costs, as observed in Experiment 2, are the cause or the consequence of attentional orienting.

It has been claimed that attentional orienting is best understood as reflecting "active sampling" within the framework of predictive processing (Dey & Gottlieb, 2019; Feldman & Friston, 2010). An implication of the predictive-processing scheme is that newly incoming data that are unlikely under some prior expectation elicit larger prediction-error responses than stimuli for which prior probability is high. Resolving these errors entails closer inspection (increased attentional sampling) of the corresponding stimuli. An important source of prior information in natural visual perception is the gist of a scene (Bar, 2004; Oliva & Torralba, 2001, 2007). Scene gist is typically defined as the holistic semantic information present in a natural image, most commonly operationalized as a category label (in our study, e.g., "kitchen" vs. "bathroom"). Observers can extract scene gist from an image very rapidly before the identification of individual objects, and it is known that scene gist influences the (later or concurrent) processing of objects within a scene (Bar, 2004; Ramkumar et al., 2016). The scene gist is naturally interpreted as inducing a prior expectation over objects to be expected within that scene. Congruent objects are easily accommodated by (i.e., a priori likely under) such a prior, whereas incongruent objects should be inspected more closely before a clear perceptual interpretation (i.e., posterior) is arrived at. This difference in amount of processing should have consequences for subjective awareness (Anzulewicz et al., 2015; Windey et al., 2013; for a schematic outlining the derivation of this hypothesis, see Fig. S13 in the Supplemental Material). The main motivation for the present study was to test the key hypothesis that gist-induced priors may lead to impaired perception of congruent items. Our results corroborate this hypothesis.

The Bayesian brain hypothesis is often invoked to explain congruency *benefits* in perception (de Lange et al., 2018), for which ample empirical evidence exists. Nevertheless, we report robust congruency *costs*. These seemingly contradictory claims are reconciled by noting that congruency benefits in scene perception are typically reported for tasks in which the gist-induced prior is, by itself, already helpful for behavior. After the gist of a scene is recognized as corresponding to a barbershop, observers will more likely identify objects inside that scene as hair dryers, even independently from the associated sensory input. Similarly, observers can use gist-based knowledge to efficiently guide a search for a hair dryer—knowledge that is unavailable when searching for a hammer in the same scene. In contrast to category identification or visual search, change detection is a nonsemantic task, in which scene gist cannot inform the judgment to be made. We deliberately designed Experiment 2 to similarly involve a judgment orthogonal to any scene-induced prior, and as predicted, this revealed congruency costs rather than benefits.

The apparent paradox between congruency benefits in some settings and congruency costs in others (or, more generally, predictive upweighting vs. cancellation) is an active area of debate and research. A recently proposed "opposing-process" model is designed to resolve this paradox by positing that perception is initially biased toward a priori likely stimuli, whereas later processes shift this balance toward (high-precision) unexpected (and thus informative) inputs (Press et al., 2020). Our observations can be neatly accommodated within this framework: High-precision sensory input corresponding to incongruent objects results in informative

prediction errors, thus "divert[ing] extra perceptual processing resources to the unexpected" (Press et al., 2020, p. 18), leading to the observed impaired perception of congruent objects.

Finally, we would like to emphasize that in addition to presenting frequentist null-hypothesis tests, we consistently tested our key hypotheses using Bayesian hierarchical models with full random effects. This allowed us to formally generalize our conclusions not just to the population from which our participants were drawn but also to the population from which our stimuli were drawn (Arnqvist, 2020; Yarkoni, 2020), lending further support to the generality of our findings.

In summary, we tested an important and seemingly counterintuitive hypothesis of the influential idea of neural predictive processing: that prior information due to real-world scene gist can lead to reduced processing of objects congruent within such scenes. Across two experiments with distinct experimental tasks, we found clear evidence in favor of this hypothesis, thereby furthering our understanding of how prior knowledge interacts with sensory input to yield real-world percepts and guide behavior.

Transparency

Action Editor: Krishnankutty Sathian
Editor: Patricia J. Bauer
Author Contributions

M. V. Peelen and F. P. de Lange contributed equally to this study. All the authors contributed to the study design. E. Spaak developed the study concept, implemented the experiments, and collected and analyzed the data. All the authors interpreted the results. E. Spaak wrote the first draft of the manuscript, and all the authors contributed to revisions and approved the final manuscript for submission.

Declaration of Conflicting Interests

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

Funding

This work was supported by The Netherlands Organisation for Scientific Research (NWO Veni Grant 016.Veni.198.065 awarded to E. Spaak and Vidi Grant 452-13-016 awarded to F. P. de Lange) and by the European Research Council under the European Union's Horizon 2020 program for research and innovation (Grant No. 725970 to M. V. Peelen and Grant No. 678286 to F. P. de Lange). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Open Practices

All data and experimental stimuli for both experiments have been made publicly available via the Donders Repository and can be accessed at https://doi.org/10.34973/x6fy-4q26. All analysis code is available from GitHub at https://github.com/Spaak/context-congruency. The design and analysis plans for the experiments were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Eelke Spaak https://orcid.org/0000-0002-2018-3364

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976211032676

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Anzulewicz, A., Asanowicz, D., Windey, B., Paulewicz, B., Wierzchoń, M., & Cleeremans, A. (2015). Does level of processing affect the transition from unconscious to conscious perception? *Consciousness and Cognition*, *36*, 1–11. https://doi.org/10.1016/j.concog.2015.05.004

Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution*, *35*(4), 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77–80.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica*, *129*(2), 255–263. https://doi.org/10.1016/j.actpsy.2008.08.006

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329. https://doi.org/10.1007/BF00868064

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, Article 548. https://doi.org/10.3389/fpsyg.2012.00548

Dey, A., & Gottlieb, J. (2019). Attention, information-seeking, and active sampling: Empirical evidence and applications

for learning. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge handbook of motivation and learning* (pp. 183–208). Cambridge University Press. https://doi.org/10.1017/9781316823279.010

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, Article 215. https://doi.org/10.3389/fnhum.2010.00215

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 760–777. https://doi.org/10.1037/0096-1523.30.4.760

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*(1–3), 213–235. https://doi.org/10.1080/135062800394775

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jeffreys, S. H. (1998). *The theory of probability* (3rd ed.). Oxford University Press.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, *23*(8), 672–685. https://doi.org/10.1016/j.tics.2019.04.013

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. A. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *The Journal of Open Source Software*, *4*(33), Article 1143. https://doi.org/10.21105/joss.01143

LaPointe, M. R. P., Lupianez, J., & Milliken, B. (2013). Context congruency effects in change detection: Opposing effects on detection and identification. *Visual Cognition*, *21*(1), 99–122. https://doi.org/10.1080/13506285.2013.787133

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565–572. https://doi.org/10.1037/0096-1523.4.4.565

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

Mack, A., Clarke, J., Erol, M., & Bert, J. (2017). Scene incongruity and attention. *Consciousness and Cognition*, *48*, 87–103. https://doi.org/10.1016/j.concog.2016.10.010

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis*. Erlbaum. http://archive.org/details/researchdesignst00jero_935

Öhlschläger, S., & Võ, M. L.-H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, *49*(5), 1780–1791. https://doi.org/10.3758/s13428-016-0820-3

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003

Ramkumar, P., Hansen, B. C., Pannasch, S., & Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study. *NeuroImage*, *134*, 295–304. https://doi.org/10.1016/j.neuroimage.2016.03.027

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, Article e55. https://doi.org/10.7717/peerj-cs.55

Schmalz, X. (2019, September 12). Bayes factors 101: Justifying prior parameters in JASP. *Xenia Schmalz's Blog*. http://xeniaschmalz.blogspot.com/2019/09/justifying-bayesian-prior-parameters-in.html

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research articles: A simplified methodology*. Work-Learning Research.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59*(11), 1931–1949. https://doi.org/10.1080/17470210500416342

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency and gist in the inspection of objects in natural scenes. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window*

on mind and brain (pp. 563–579, IV–VII). Elsevier. https://doi.org/10.1016/B978-008044980-7/50028-8

Vallat, R. (2018). Pingouin: Statistics in Python. *The Journal of Open Source Software*, *3*(31), Article 1026. https://doi.org/10.21105/joss.01026

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Van Rossum, G., & Drake, F. L., Jr. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3), Article 24. https://doi.org/10.1167/9.3.24

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., . . . Brian. (2020). *mwaskom/seaborn: V0.10.1 (April 2020)*. Zenodo. https://doi.org/10.5281/zenodo.3767070

Windey, B., Gevers, W., & Cleeremans, A. (2013). Subjective visibility depends on level of processing. *Cognition*, *129*(2), 404–409. https://doi.org/10.1016/j.cognition.2013.07.012

Yarkoni, T. (2020). The generalizability crisis. *Behavioral & Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2016). *Bambi: A simple interface for fitting Bayesian mixed effects models*. https://doi.org/10.31219/osf.io/rv7sn

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica y de Investigacion Operativa*, *31*(1), 585–603. https://doi.org/10.1007/BF02888369