Supplementary Material

XGE-2022-0160

1. Task × Trial Type Interaction

1.1. Reaction Times

An overview of the observed Stroop effects by task and experiment and of reaction times by trial type, task, and experiment are displayed in Figures S1 and S2, respectively. In all three experiments, the repeated-measures ANOVA reveals a main effect of trial type (congruent vs. incongruent: Experiment 1: $F(1, 49) = 28.689$, $p < .001$, Experiment 2: $F(1, 49) = 6.18$, $p = .016$; Experiment 3: $F(1, 49) = 60.904$, $p < .001$), a main effect of task (*Which one is larger on the screen?* vs. *Which one is smaller on the screen?*: Experiment 1: $F(1, 49) = 26.815$, $p < .001$, Experiment 2: $F(1, 49) = 6.989$, $p = .011$; Experiment 3: $F(1, 49) = 23.985$, $p < .001$), and, with the exception of Experiment 2, a trial type × task interaction (Experiment 1: $F(1, 49) = 29.123$, $p < .001$, Experiment 2: $F(1, 49) = 3.646$, $p = .062$; Experiment 3: $F(1, 49) = 16.325$, $p = .002$). The main effect of task arises because in all three experiments, participants are quicker to solve the *longer* task than the *smaller* task, on average. The interaction arises because in all three experiments, the Stroop effect is stronger in the *smaller* task.

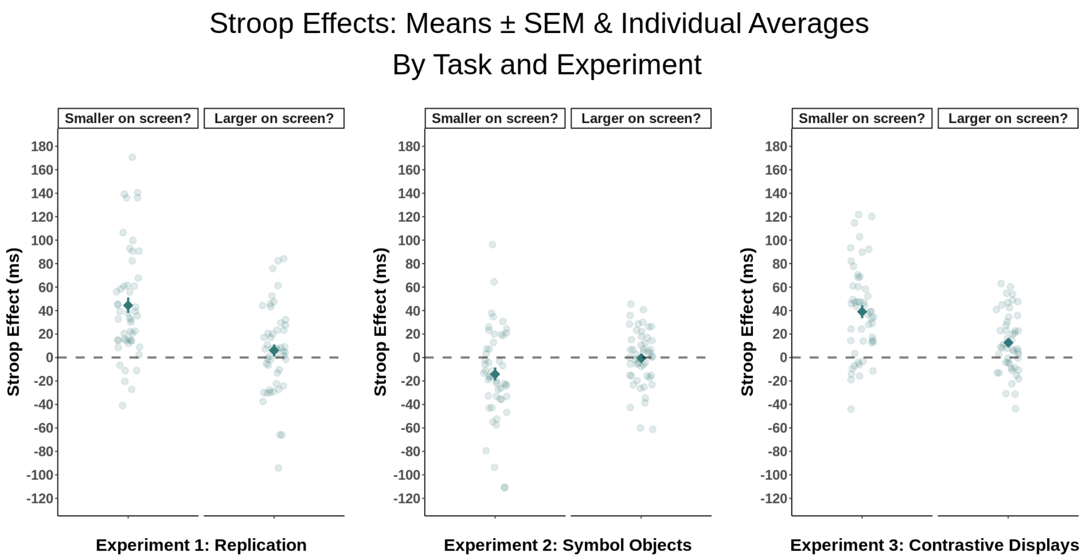


Figure S1. Stroop Effects in Experiments 1-3 by task. Transparent circles represent individual Stroop effects (incongruent – congruent reaction times); opaque diamonds show group average Stroop effect ± 1 SEM.

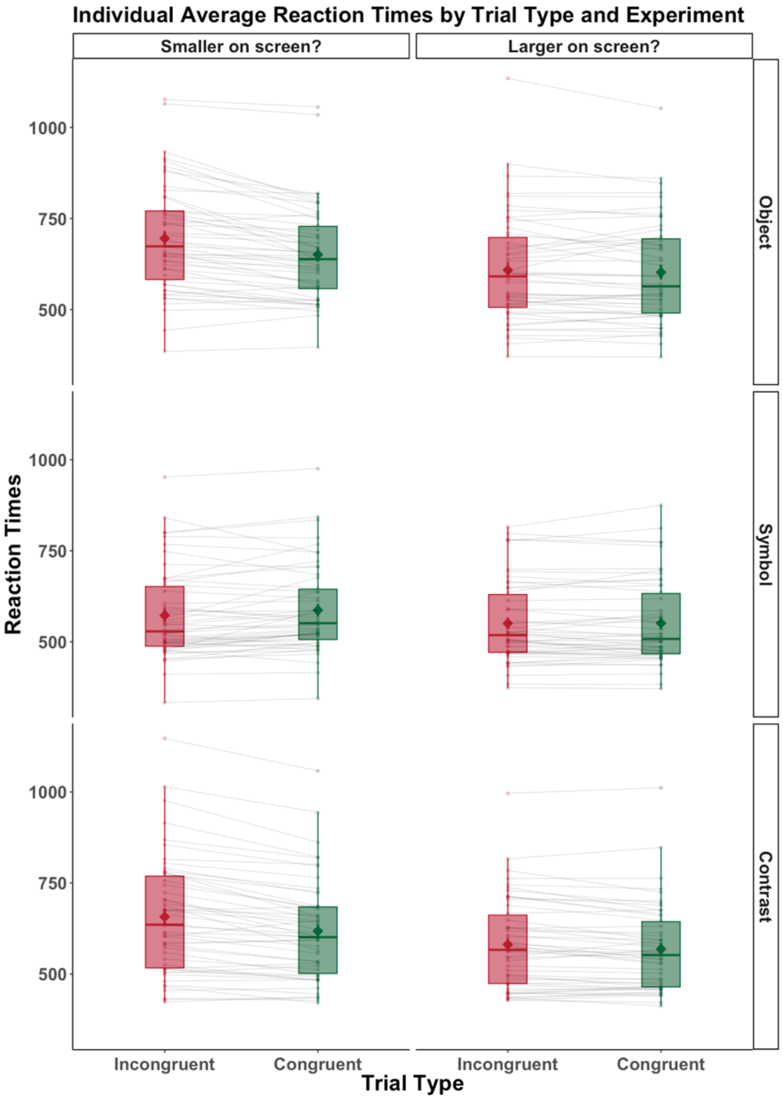


Figure S2. Reaction times in Experiments 1-3 by trial type and task. Transparent circles and the lines connecting them represent individual reaction times as a function of trial type; opaque diamonds depict group averages ± 1 SEM; boxplots indicate the median and interquartile range.

In Experiment 1, if we look at reaction times data separately by task, the Stroop effect turns out to have been driven mostly by the *smaller*-task judgments, $t(49) = 6.717$, $p < .001$, Cohen's $d = 0.95$, 95% CI [0.618, 1.294], as there was little difference between incongruent and congruent trials in the *larger* task, $t(49) = 1.174$, $p = .246$, Cohen's $d = 0.167$, 95% CI [–0.115, 0.449]. This partly replicates the original finding, as Konkle & Oliva (2012) also found a stronger effect in the *smaller* task.

In Experiment 2, the effect was also primarily driven by the *smaller* task, $t(49) = –2.524$, $p = .015$, Cohen's $d = 0.357$, 95% CI [0.07, 0.648]; in the larger task, trial type had no effect on reaction times, $t(49) = –0.189$, $p = .851$, Cohen's $d = 0.027$, 95% CI [–0.253, 0.307]).

In Experiment 3, we also found a larger incongruent-congruent difference in the *smaller*-judgment task, $t(50) = 7.091$, $p < .001$, Cohen's $d = 0.963$, 95% CI [0.665, 1.354]. However, unlike in Experiments 1 and 2, we also find a Stroop effect in the *larger*-judgment block, $t(50) = 3.506$, $p < .001$, Cohen's $d = 0.496$, 95% CI [0.202, 0.796].

1.2. Errors

An overview of participants' error rates by task and trial type in all three experiments is displayed in Figure S3. In Experiment 1, a repeated-measures ANOVA reveals a main effect of trial type, $F(1, 49) = 28.301$, $p < .001$, no main effect of task, $F(1, 49) = 2.054$, $p = .158$, and a trial type × task interaction, $F(1, 49) = 6.462$, $p = .014$. If we split the results by task, we find that in Experiment 1, participants made many more errors on incongruent trials in the *smaller* task, $t(49) = 5.48$, $p < .001$, while this difference was less pronounced in the *larger* task, $t(49) = 2.11$, $p = .040$.

In Experiment 2, a repeated-measures ANOVA reveals a main effect of trial type, $F(1, 49) = 17.928$, $p < .001$, a main effect of task, $F(1, 49) = 18.081$, $p < .001$, and a trial type × task interaction, $F(1, 49) = 14.032$, $p < .001$. If we split the error rates by task, we find the same pattern on the *smaller* task as in Experiment 1, except in the opposite direction, $t(49) = –5.216$, $p < .001$, and no significant effect in the larger task even though the difference was in the same direction as that of the *smaller* task, $t(49) = –1.218$, $p = .229$.

In Experiment 3, a repeated-measures ANOVA indicates a main effect of trial type, $F(1, 49) = 13.920$, $p < .001$, a main effect of task, $F(1, 49) = 5.833$, $p = .02$, and no trial type × task interaction, $F(1, 49) = .01$, $p = .922$.

If we split the error rates by task, the pattern of Experiments 1 and 2 reverses. The incongruent-congruent difference in error rates is narrower on the *smaller* task, *t*(49) = 2.055, *p* = .045, than on the larger task, *t*(49) = –3.58, *p* = .001.

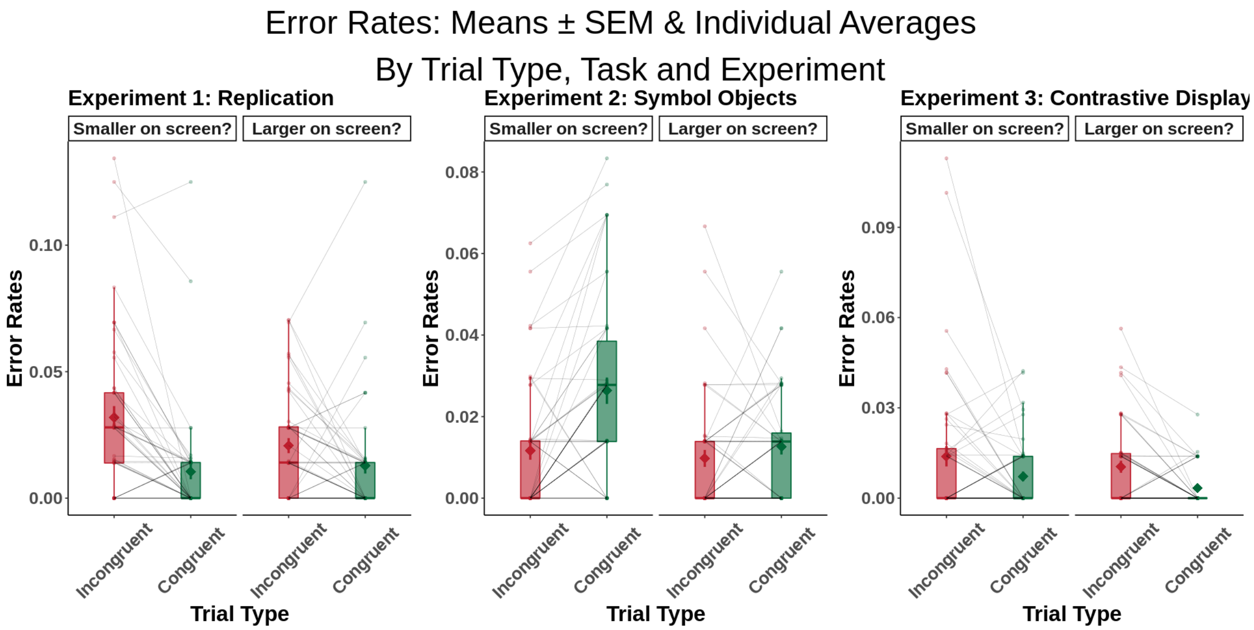


Figure S3. Error rates in Experiments 1-3 by trial type and task. Transparent circles and the lines connecting them represent individual error rates as a function of trial type; opaque diamonds depict group averages ± 1 SEM; boxplots indicate the median and interquartile range.

2. Item Effects

In each experiment, for each pair, and for each participant, we obtain a Stroop effect by averaging over task (*smaller* vs. *larger*) and side of presentation (*left* vs. *right*). In Figure S4, we plot the pair Stroop effects sorted in decreasing order of their mean, together with error bars representing ±1SEM.

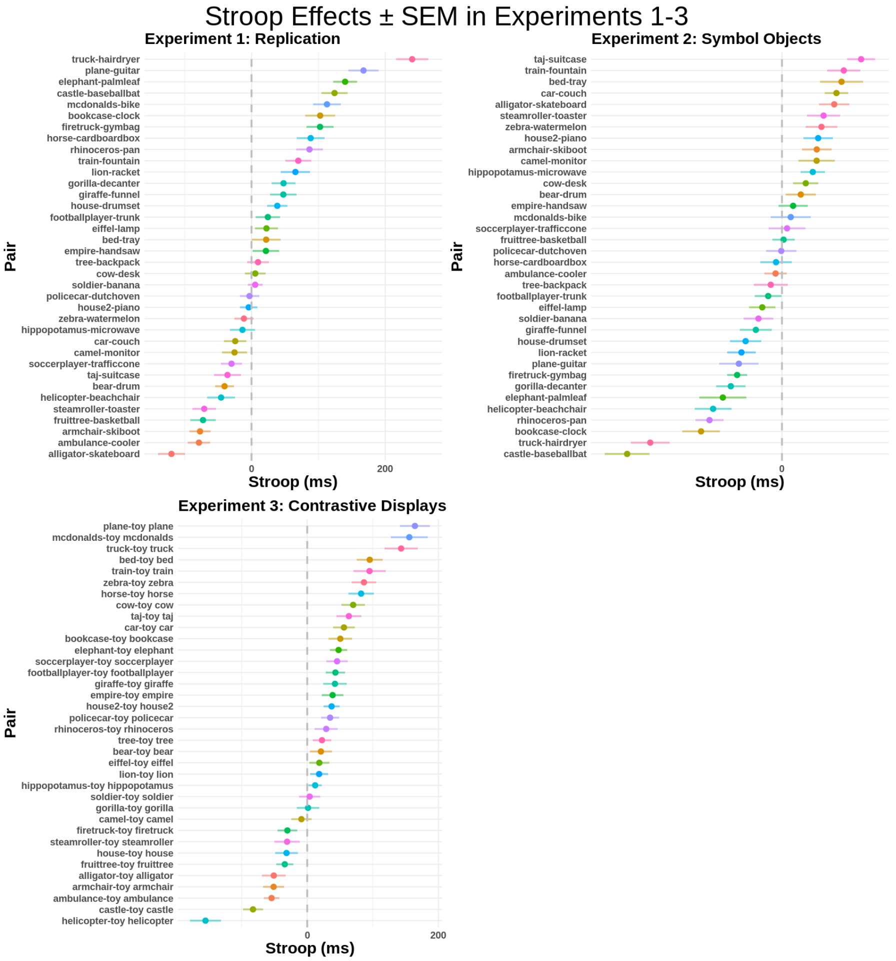


Figure S4. Item effects in Experiments 1-3 sorted by the Stroop effect in descending order. Circles represent averages over participants; lines represent ±SEM.

3. Trial type × Animacy Interaction

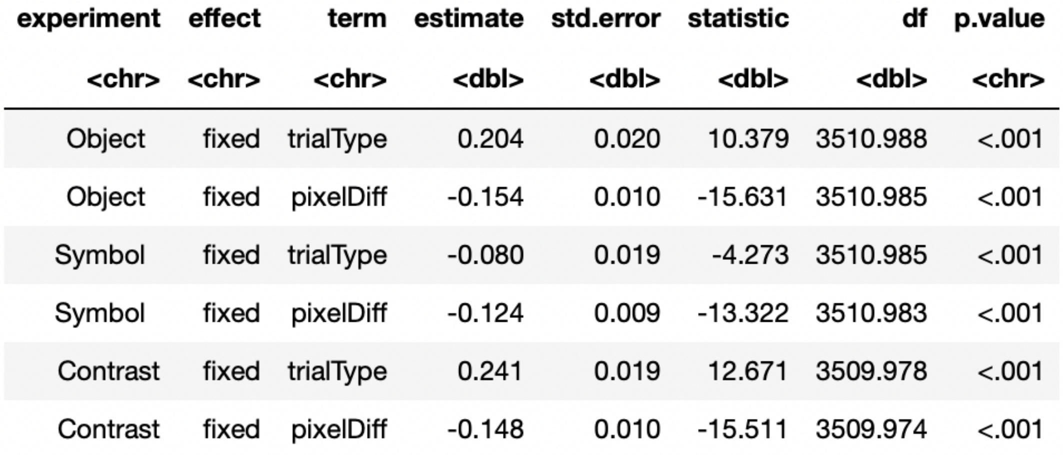
In each of the three experiments, 15 out of the 36 pairs contained one image of an animate entity (e.g., camel-monitor; toy zebra-watermelon—in Experiment 2, we coded toy animals as animate). If we test for a trial type × animacy interaction, we find that animacy does not significantly interact with trial type in terms of reaction times in any of the three experiments: Experiment 1, $F(1, 49) = 1.81$, $p = .185$; Experiment 2: $F(1, 49) = .405$, $p = .527$; Experiment 3: $F(1, 49) = 1.02$, $p = .318$.

4. Pixel Area Differences Control

If the larger objects in our stimuli filled more of their bounding box than the smaller objects, it could have introduced a bias in the task. In that case, congruent trials would become easier (on congruent trials, the larger-in-the-world object appears larger because it fills more of its bounding box, and therefore it becomes easy to make the visual judgment task) and incongruent trials would become harder (on incongruent trials, the larger-in-the-world object is depicted at a small size but it appears larger because, again, it fills more of its bounding box, leading to a more difficult judgment), resulting in a Stroop effect even if participants do not compute real-world sizes at all.

To control for such a bias, we obtained the ratio of non-white to white pixels for each of the 108 images in the stimuli set (36 triplets of 3 images each). Then, for each pair, we obtained the difference between the white-to-non-white pixel ratio of both images. The means of the pixel area differences between the paired objects in the three experiments were close to 0 and to each other (Experiment 1: –0.04; Experiment 2: 0.022; Experiment 3: –0.018). Statistical tests confirmed that the null hypotheses for the means being equal to 0 (all *p*-values > .16) and for the means being equal to each other (*p* = .309) cannot be rejected. This implies that the pixel area differences have likely increased the noise in our measurement without biasing it.

To confirm this, we recoded pixel difference to measure the difference between the pixel area filled by the larger image (not necessarily the larger depicted object) and the smaller image on a trial-by-trial basis. We averaged reaction times over participants, item pairs, and trial type, and scaled reaction times and pixel area differences by dividing each by their standard deviations. We then fitted a linear mixed model for each experiment, with trial type and pixel area differences as fixed effects, and participant ID and item pair as random effects. In all three experiments, trial type remains a significant predictor of reaction times when pixel area differences are controlled for (Table S1).



experiment	effect	term	estimate	std.error	statistic	df	p.value
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Object	fixed	trialType	0.204	0.020	10.379	3510.988	<.001
Object	fixed	pixelDiff	-0.154	0.010	-15.631	3510.985	<.001
Symbol	fixed	trialType	-0.080	0.019	-4.273	3510.985	<.001
Symbol	fixed	pixelDiff	-0.124	0.009	-13.322	3510.983	<.001
Contrast	fixed	trialType	0.241	0.019	12.671	3509.978	<.001
Contrast	fixed	pixelDiff	-0.148	0.010	-15.511	3509.974	<.001

Table S1. Subset of the output of linear mixed model, *response time ~ trialType + pixelDiff* + (1|*pair*) + (1|*ID*), by experiment.

Pixel area differences contributed significantly to the Stroop effect in all three experiments, as indicated by the fact that the *pixelDiff* coefficients were significantly below 0. This is to be expected: irrespective of object size measurement and trial congruency, if the larger image fills more of its bounding box than the smaller image, it makes image size judgment easier and reaction times shorter. However, trial type continued to be a significant predictor even after controlling for pixel area differences, suggesting that object size inferences were also made. Note also that while the items in Experiment 3 are better matched in terms of pixel area, the effect of pixel area differences on reaction times was similar to those in Experiments 1 and 2.

5. Correlations between Size Disparity and Stroop Effect by Item Pair

Size disparity between the depicted object might also have contributed to the Stroop effect found in Experiments 1-3. We obtained the size of the real-world referent depicted in all 108 images based on the procedure in Konkle & Oliva (2011), either from https://konklab.fas.harvard.edu (for the stimuli which were used in both Konkle & Oliva, 2011 and in our study) or by searching the internet for the typical dimensions of the depicted object. Following Konkle & Oliva (2011), we took the logarithm of the diagonal of the 2-D bounding rectangle of the object, ignoring depth. For each trial, we obtained the real-world size difference by subtracting the logarithm of the large image referent size from the logarithm of the small image referent size. In Experiment 2, we used both the real-world size of the small toy and of its referent in separate analyses. Figure S5 depicts the measured Stroop effects, averaged by pair, against the log-size difference between the large object and the small object in each pair.

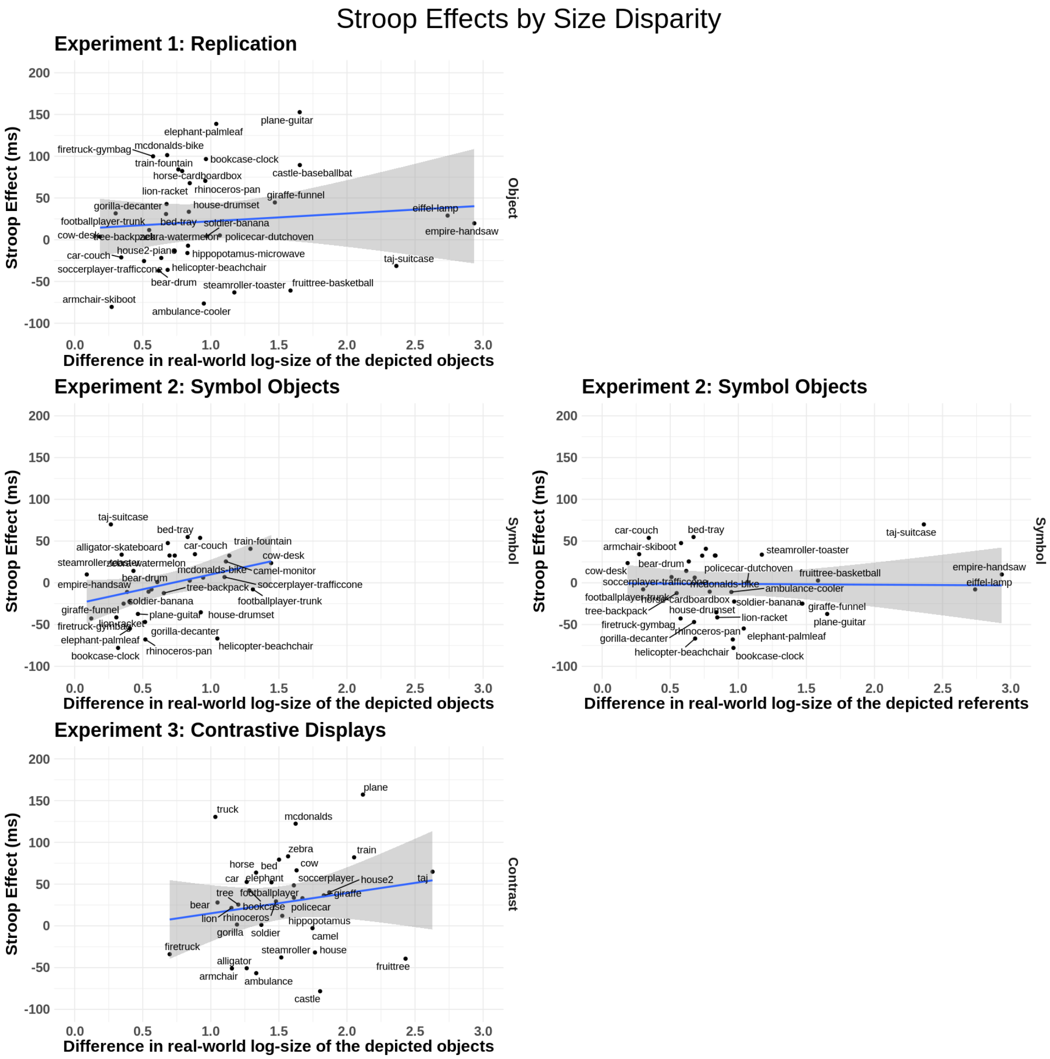


Figure S4. (Left) Stroop effect (incongruent – congruent) by size disparity in Experiments 1-3. (Right) Stroop effect (congruent – incongruent) by size disparity of the toys' referents in Experiment 2. The grey-shaded area represents the 95% confidence interval around the regression line.

We found no effect of the magnitude of size difference on Stroop effects in Experiments 1 and 3 (*p*s > .49), suggesting that, in general, the Stroop delay was not affected by the magnitude of the perceived incongruency (Figure S5, left column). In Experiment 2, we found a positive correlation between the difference in log-size and the Stroop effect (*p* = .004) when the size disparity was measured against the toy size. This indicates that the higher the difference between a toy and a medium-sized object was, the more likely it was that participants perceived the trials in which the toy was depicted larger as incongruent. Note, however, that the regression line for this experiment is not analogous to the ones for Experiments 1 and 3. There, the fitted lines are projected towards 0 when the log-size goes to 0, as expected (indicating that the Stroop effect disappears when there is no difference between the actual sizes of the objects). In Experiment 2, when the correlation is calculated with the size measurement of the toys (left column), the regression line predicts a negative Stroop effect when there is no difference between the two objects, which makes no sense. However, if the participants interpreted toy images as standing for the referents of the toys, the incongruent trials (in which the toy was displayed larger than the actually larger object) were actually easier and thus more congruent for participants. Indeed, if the regression is calculated with the size measurements of the toys' referents, rather than with the size measurements of the toys (Figure S5, right column), the regression line looks identical to to those of Experiments 1 and 3: a non-significant correlation (p = .587), as well as a plausible prediction when the difference goes to 0.

References

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37. https://doi.org/10.1037/a0020413

Konkle, T., & Oliva, A. (2012). A familiar-size Stroop effect: real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561. https://doi.org/10.1037/a0028294