SUPPLEMENTAL ONLINE MATERIAL

1. Possible Explanations of Petrova et al.'s Failures to Replicate

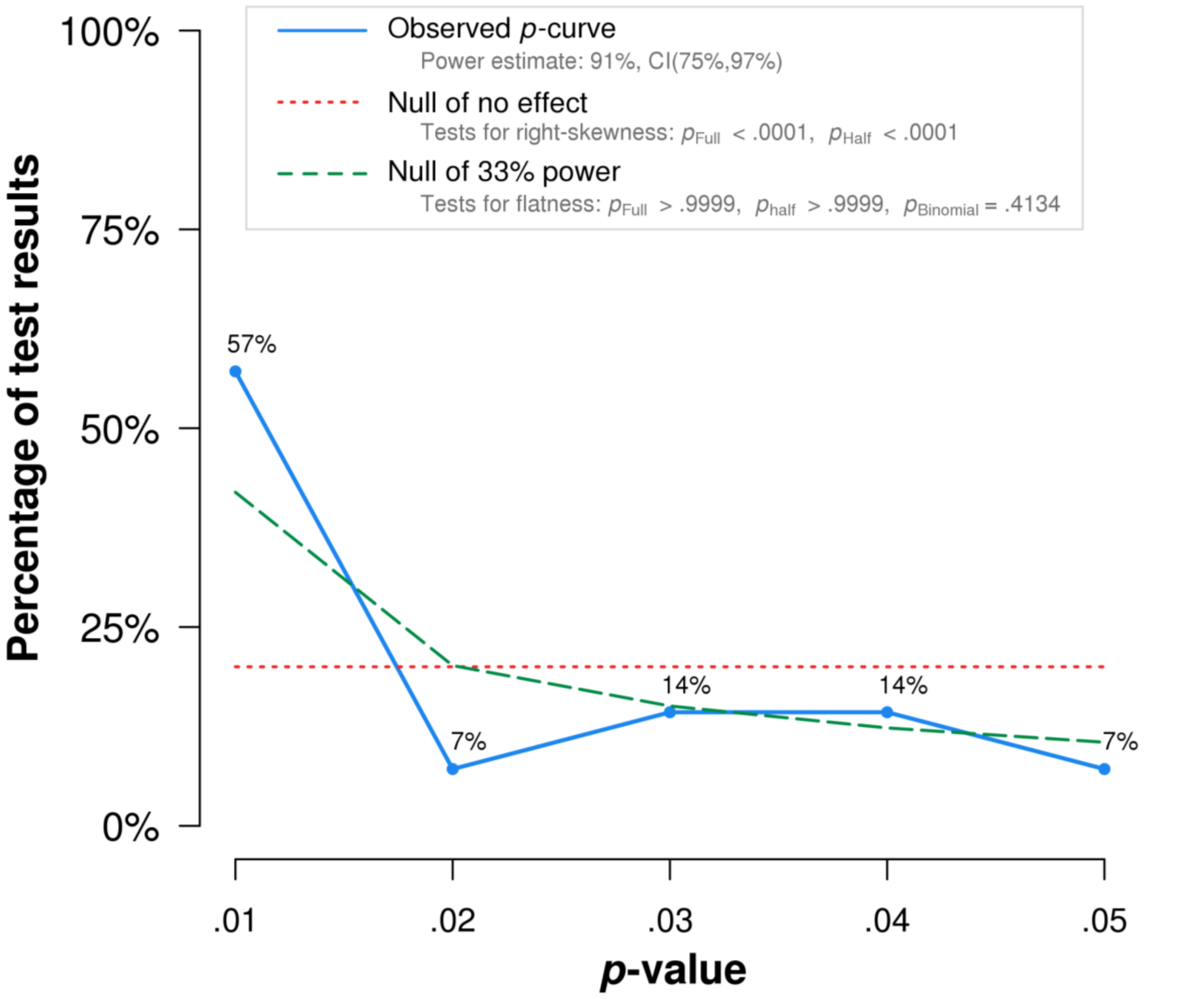
Failures to replicate, like those reported by Petrova et al. (2018), may arise from five potential causes. As described in more detail below, the first three explanations attribute failure to the original study, questioning whether the effect is real (as in Petrova et al.). The other two potential explanations instead attribute failure to the replication study, questioning whether the replication attempt is of sufficient quality and similarity to convincingly demonstrate that the original effect is not real. Each of these five explanations is addressed in detail next.

(1) The original result is a false-positive, Type I error. In the case of spatial interference, this explanation is extremely unlikely. Fourteen significant demonstrations of the effect have been published. If these were false-positives, approximately 560 tests of the effect must have been conducted.[1] This is an implausibly large "file drawer."

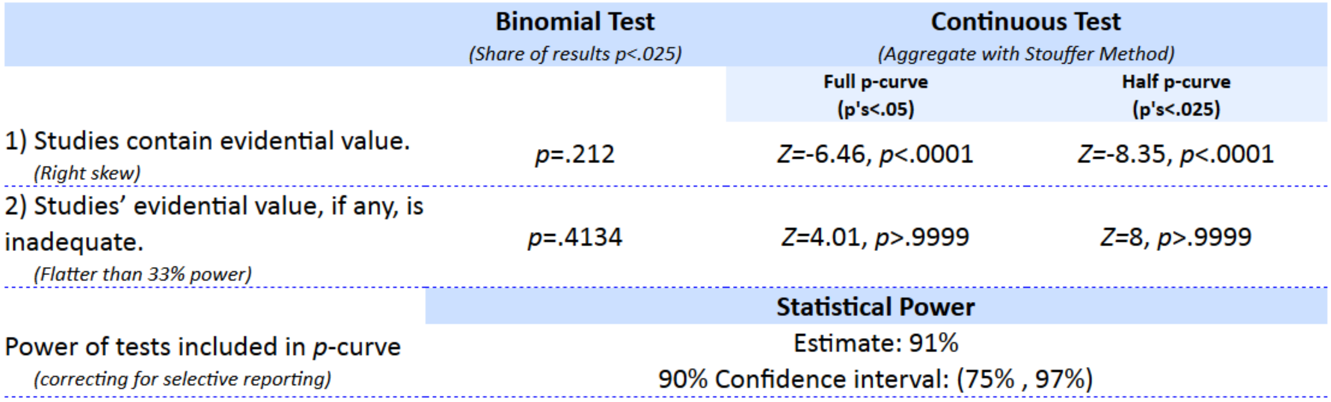
(2) The original result was obtained by data manipulation (i.e., *p*-hacking; Simmons, Nelson, & Simonsohn, 2011). To test the plausibility of this explanation, we submitted the 14 significant demonstrations of the spatial interference effect (Table 1) to a *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2014). Results are shown in Figure S1. Collectively these results provide positive evidence of the effect, $Z = -6.46$, $p < .0001$, with an estimated power of 91%. Thus, there is no evidence that significant demonstrations of the effect were *p*-hacked.

(3) The original result was obtained by data fabrication (i.e., outright fraud). Given 14 significant demonstrations of the effect by 4 independent research groups, this is not a viable explanation of the spatial interference effect.

[1] The spatial interference effect is a directional hypothesis. So whereas the probability of obtaining a false-positive *in either direction* is $p = .05$, the probability of obtaining a false-positive *interference* effect is $p = .025$. A false-positive interference effect is therefore expected about once out of every 40 tests. In order to produce 14 false-positive interference effects, about 14 × 40 = 560 experiments would need to be conducted.



Note: The observed *p*-curve includes 14 statistically significant (*p* < .05) results, of which 9 are *p* < .025. There were no non-significant results entered.



	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve **(p's<.05)**	**Half p-curve** **(p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.212	*Z*=-6.46, *p*<.0001	*Z*=-8.35, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*=.4134	*Z*=4.01, *p*>.9999	*Z*=8, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 91% 90% Confidence interval: (75% , 97%)		

Figure S1. *P*-curve of 14 significant demonstrations of the spatial interference effect. Evidence of the spatial interference effect was strong and significant; there was no evidence of *p*-hacking.

If reported spatial interference effects are not false-positives, nor *p*-hacked or fraudulent, they are likely to be real effects. So how can researchers (e.g., Petrova et al.) fail to replicate them? The two remaining potential explanations address this question.

(4) The replication result is a false-negative, Type II error. Of Petrova et al.'s 16 tests of the effect, 10 were conducted under suitable conditions (i.e., SOA < 400 ms; see Table 1) and with methods and analyses that appear sound and rigorous. Yet they obtained significant spatial interference only once. A serious methodological limitation of Petrova et al.'s replication studies, and one that contributes substantially to false-negatives, is that each of their studies was underpowered (see Simonsohn, 2015).[2] Nonetheless, when all of Petrova et al.'s studies are combined, the collective sample size becomes quite large (total $N = 265$), and still their collective effect is nonsignificant. Thus, we doubt that Petrova et al.'s failure to replicate is a false-negative.

(5) The effect is moderated by a variable (or variables) on which the original and replication studies differ. Given the presence of salient differences between the original and replication studies (e.g., English v. Italian language), this explanation of Petrova et al.'s failure to replicate seems most plausible.

We therefore sought to identify potentially important differences between the successful demonstrations of the effect and the failures to obtain the effect. Specifically, we sought to identify moderators via meta-analysis (Braver et al., 2014). We began with a relatively inclusive sample of tests of the effect, and then we progressively restricted the sample by applying moderators, until the final sample included only tests that closely resembled the original study (Estes et al., 2008). The details of

[2] To achieve sufficient statistical power (about 80%) to detect the original effect, replication studies should use samples at least 2.5 times larger than the original (Simonsohn, 2015). In the present case, the original study (Estes et al., 2008, Experiment 3) had 30 participants, so each of Petrova et al.'s replication attempts should have included at least 75 participants. Instead, their sample sizes ranged from 18 to 41, with an average sample size ($N = 27$) that was actually smaller than the original study. Thus Petrova et al.'s replication studies were substantially underpowered.

this funneled moderator analysis follow in Section 3 below. Before describing our own selection criteria, however, we first consider Petrova et al.'s selection criteria.

2. Petrova et al.'s Selection Criteria

Petrova et al. define the effect as follows: "Estes, Verges and Barsalou (2008) reported that reading a word with a spatial connotation (sky) interfered with the subsequent identification of an *unrelated visual stimulus* (letter X or O) presented in a semantically related portion of the screen (location cue congruency, LLC effect)" (p. 2). They then go on to explain that "…Estes et al.'s results (Experiment 3) are particularly relevant for several reasons…interference was observed despite the words being presented in isolation, with *no preceding context*, *no task to be performed on them*, and *short delays between the cue word and the target stimulus*. Therefore, given that the effect was obtained *without explicit or implicit reference to the spatial properties of the words*, and that there was no benefit to processing them, the interference effect suggests that spatial information is mandatorily and rapidly activated during language processing" (p. 2). In these two descriptions Petrova et al. state five main selection criteria for their investigation: (1) no reference to the spatial properties of the words; (2) no semantic context; (3) no task performed on the cue words; (4) short SOA; (5) semantically unrelated target. We have italicized the relevant parts of the quotes above to indicate where these five characteristics are evident.

Petrova et al. did not consistently apply these criteria. Table S1 shows for each of their nine experiments which of these five criteria they applied. As evident in the table, across different experiments they violated three of their five stated criteria of investigation. Furthermore, Petrova et al. included in their meta-analysis another experiment that violated these criteria (i.e., in Gozli et al., 2013, Experiment 3B used a long SOA).

Table S1. Petrova et al.'s stated selection criteria, and evaluation of their nine experiments in terms of those criteria.

Exp	No Reference to Spatial Associations	No Semantic Context	No Cue Judgment	Short SOA	Unrelated Target
1	√	√	√	X	√
2	√	√	X	X	√
3	√	√	X	X	√
4	√	√	√	√	√
5	√	√	√	√	√
6	√	√	√	√	√
7	X	√	√	√	√
8	X	√	√	√	√
9	√	√	√	√	√

These criteria violations created difficulty for us in terms of how to define the selection criteria of our own meta-analysis. If we had adopted their stated criteria, then we would have had to exclude several of their experiments. We therefore instead adopted our own, theoretically motivated criteria that were similar but not identical to Petrova et al.'s stated criteria.

3. Funneled Moderator Analysis (Meta-Analysis)

3.1. Methods

Sampling. Our general purpose was to conduct a relatively comprehensive review and test of the spatial interference effect. Our sampling procedure therefore used the following inclusion criteria.

1. Unlike Petrova et al., we included studies in which the linguistic cues were single words, pairs of words, or minimal sentences. Bergen et al. (2007) used very brief sentence cues (e.g., "The patient rose"), and Estes et al. (2008) used word pairs (e.g., "cowboy hat") in their Experiments 1 and 2. We believe that including those initial studies of spatial interference, with minimal semantic contexts, is important for adequately characterizing the spatial interference effect. Their inclusion also allows us to test whether the presence of a semantic context moderates the spatial interference effect.

2. We included only studies in which the linguistic cues were from multiple categories (e.g., clothing, animals, body parts, etc.). This excludes studies in which the cues were all from the same category (e.g., Gozli et al., 2013, Experiments 1, 2, and 5). Single-category cues allow participants to establish a spatial reference frame that remains constant across the experiment, creating a more consistent stimulus-response mapping across trials. This consistent mapping renders responding more efficient, thereby eliminating the interference effect and sometimes producing facilitation instead (Gozli et al., 2013; Ostarek & Vigliocco, 2017).

3. Following Petrova et al. (2018), we included only studies in which the visual target was abstract and unrelated to the linguistic cue (e.g., X or O, ■ or ●). This excludes studies in which the visual target was a familiar object (Estes et al., 2015, Experiments 1 and 2; Ostarek & Vigliocco, 2017), because familiar objects have their own spatial and semantic associations that affect responding to targets at different locations (Ostarek & Vigliocco, 2017).

4. Following Petrova et al. (2018), we included only studies in which the task was to *identify* the visual target. This excludes studies using a target *detection* task, which does not require detailed perceptual processing of the visual target and therefore typically exhibits facilitation rather than interference (Gozli et al., 2013).

5. Individual experiments that included multiple tests of the hypothesis were treated as separate tests. For instance, an experiment with two SOA conditions was treated as two tests of spatial interference.

6. We excluded experimental conditions that were specifically intended to eliminate the spatial interference effect. For instance, Estes et al. (2008, Experiment 2) included a “masked” condition that was intended to block spatial interference, so that condition was excluded and only the “unmasked” condition was included.

Based on these criteria, the final sample included 37 tests of the spatial interference effect, listed in Table 1. These tests are distributed across 6 published papers and 3 unpublished studies by 5 independent research groups. Two of those 37 tests require further clarification: (1) Gozli et al. (2013) discovered a problematic cue category in their Experiment 6 (“power” cues), and so reported analyses without those stimuli. We therefore also sampled their result with these problematic stimuli excluded. (2) Estes et al. (2015, Experiments 3 and 4) analyzed spatial associations of cue words as a continuous factor. Although more sensitive, their approach prevents direct comparison with the categorical (upward vs. downward cues) approach used in all other tests of the effect. We therefore reanalyzed those data in a categorical manner, using the cues’ classification in prior studies as upward or downward, excluding those that in prior studies were non-spatial filler cues (cf. Estes et al., 2008).

Effect size. Given that all tests of this effect used the same, directly interpretable dependent variable (i.e., response times), we report raw effect sizes (milliseconds, *ms*; Bond, Wiitala, & Richard, 2003). Effect size was calculated simply as $M_{\text{congruent}} – M_{\text{incongruent}}$, where “Congruent” and “Incongruent” respectively refer to trials in which the visual target appears in the location associated with the cue word (e.g., “bird” → target at top) or in the opposite location (e.g., “bird” → target at bottom). Effect sizes and 95% *CI*s were calculated directly from the raw data when possible, or were estimated from descriptive statistics when raw data were not available. Effect sizes of individual tests are reported in Table 1 and illustrated in Figure 1. For comparison, the classic semantic priming effect is about 26 ms (Hutchison et al., 2013).

Meta-analyses. Fixed-effects models test the reliability of an effect among the previously observed data, whereas random-effects models test whether a presumed effect is likely to generalize beyond the observed data. In the context of a replication study (as in Petrova et al., 2018), researchers may be interested in both questions: Is the effect reliable among prior studies, and is it likely to generalize to new studies? We therefore conducted two separate analyses, one with fixed-effects and one with random-effects. We first calculated the meta-analytic effect sizes and statistical comparisons using a fixed-effects model, to test whether there is reliable evidence of spatial interference in the prior studies. This analysis is also reported in the main text. We then replicated those analyses, but with a maximum likelihood random-effects model, to test whether the spatial interference effect is likely to generalize beyond the current studies. Analyses were conducted in SPSS using macros by Wilson (2006).

3.2. Results

Fixed-effects model. Across the 37 tests, the spatial interference effect was small (M = 4.52 ms, SE = .79, 95% CI [2.97, 6.07]) but significant, Z = 5.71, p < .001. There was also significant heterogeneity among the individual effects, Q(36) = 273.13, p < .001, suggesting the utility of a moderator analysis (Braver et al., 2014). Because spatial interference is known not to occur at longer SOAs (i.e., those greater than 400 ms; Gozli et al., 2013), we first tested SOA as a presumed moderator. Indeed, SOA significantly moderated the effect, Q(1) = 42.97, p < .001. The 28 tests with short SOAs (i.e., 400 ms or less) exhibited significant spatial interference (M = 8.16, SE = .97, CI [6.26, 10.05], Z = 8.44, p < .001), whereas the 9 tests with long SOAs instead exhibited modest but significant spatial facilitation (M = -2.87, SE = 1.38, CI [-5.57, -0.17], Z = -2.08, p = .037). At longer SOAs, the perceptual simulation of the linguistic cue (e.g., “bird”) dissipates, leaving one’s visual attention in the cued location without perceptual competition, producing facilitation instead of interference (Gozli et al., 2013). Thus, as expected, SOA moderates the spatial interference effect (Gozli et al., 2013). Studies with long SOAs do not provide felicitous tests of spatial interference. We therefore included only tests with short SOAs in the subsequent analyses.

Next, because spatial interference requires semantic processing of cue words (Lebois et al., 2015), and because semantic processing of cue words is more likely in orthographically deep languages (e.g., English) than in orthographically shallow languages (e.g., Italian and German; Bates et al., 2001;

Schmalz et al., 2015), we tested orthographic depth as a potential moderator. Indeed, orthographic depth significantly moderated the effect, *Q*(1) = 85.51, *p* < .001. Spatial interference was significant among the 16 tests in English with short SOAs (*M* = 19.27, *SE* = 1.54, *CI* [16.24, 22.29], *Z* = 12.50, *p* < .001), but not among the 12 tests in Italian or German with short SOAs (*M* = .96, *SE* = 1.24, *CI* [-1.47, 3.40], *Z* = .78, *p* = .44). Thus, orthographic depth appears to moderate the spatial interference effect via semantic processing. Given its post hoc nature, however, this observation warrants further investigation. Please see the main text for a brief review of relevant literature.

Finally, because richer semantic contexts may evoke stronger perceptual simulations (Wilson-Mendenhall et al., 2013) and hence larger spatial interference effects, we tested the presence of semantic context (i.e., brief sentences in Bergen et al., 2007; word pairs in Estes et al., 2008) as a potential moderator. Indeed, among the 16 tests in English with short SOAs, spatial interference was twice as large with semantic context (*N* = 6, *M* = 34.51, *SE* = 4.61, *CI* [25.47, 43.56], *Z* = 7.48, *p* < .001) as without it (*N* = 10, *M* = 17.35, *SE* = 1.64, *CI* [14.14, 20.55], *Z* = 10.60, *p* < .001). This moderation was significant, *Q*(1) = 12.30, *p* < .001.

Random-effects model. Across all tests, the spatial interference effect was small (*M* = 8.98 ms, *SE* = 2.33, 95% *CI* [4.41, 13.54],) but significant, *Z* = 3.85, *p* < .001, with significant heterogeneity among the individual effects, *Q*(36) = 273.13, *p* < .001. SOA significantly moderated the effect, *Q*(1) = 7.59, *p* = .006. Spatial interference was significant among the 28 tests with short SOAs (*M* = 13.92, *SE* = 3.13, *CI* [7.79, 20.05], *Z* = 4.45, *p* < .001), but not among the 9 tests with long SOAs (*M* = -2.58, *SE* = 5.11, *CI* [-12.59, 7.43], *Z* = -.51, *p* = .61). Orthographic depth also significantly moderated the effect, *Q*(1) = 27.51, *p* < .001. Spatial interference was significant among the 16 tests in English with short SOAs (*M* = 26.41, *SE* = 3.56, *CI* [19.45, 33.38], *Z* = 7.43, *p* < .001), but not among the 12 tests in Italian or German with short SOAs (*M* = .49, *SE* = 3.43, *CI* [-6.25, 7.22], *Z* = .14, *p* = .89). Finally, although spatial interference was larger with semantic context (*N* = 6, *M* = 32.57, *SE* = 9.32, *CI* [14.30, 50.83], *Z* = 3.50, *p* < .001) than without it (*N* = 10, *M* = 25.46, *SE* = 5.75, *CI* [14.20, 36.72], *Z* = 4.43, *p* < .001), this moderation was not significant, *Q*(1) = .42, *p* = .52.

In sum, the random-effects model largely reproduced the results of the fixed-effects model, with a significant overall effect and significant moderation by SOA and orthographic depth. This indicates that the effects observed in prior studies are likely to generalize to new studies with similar methods. However, the significant moderation by semantic context observed in the fixed-effects model was not significant here in the random-effects model, perhaps due to a lack of power (i.e., too little evidence to generalize beyond the prior studies). In fact, no prior study has directly compared the spatial interference effect with and without a semantic context. The preliminary evidence here suggests that this may be a fruitful direction for further research.

Robustness against dependency. The meta-analyses reported above assume that all 37 tests of the spatial interference effect are independent. This assumption is valid among different experiments, and among different conditions within a given experiment so long as the independent variable is manipulated between-participants. In their Experiments 1-3, however, Petrova et al. (2018) manipulated SOA within-participants. Consequently, the effect sizes among the different SOA conditions within each of those experiments are theoretically dependent. Therefore, to test whether the dependence among effects in those three experiments affected the overall pattern of results reported above, we replicated the preceding analyses (fixed-effects model), but including only the short SOA conditions of Petrova et al.'s Experiments 1-3. That is, we excluded the 450 and 900 ms delay conditions of those three experiments, leaving 31 tests of the spatial interference effect. The pattern of significant results remained unchanged. Overall effect: *Z* = 6.34, *p* < .001. Moderation by SOA: *Q*(1) = 38.91, *p* < .001. Moderation by orthography: *Q*(1) = 85.51, *p* < .001. Moderation by semantic context: *Q*(1) = 12.30, *p* < .001. Thus, the observed results were unaffected by statistical dependency among Petrova et al.'s Experiments 1-3.

3.3. Conclusion

The spatial interference effect varies systematically across task contexts. Spatial interference does not occur at long SOAs (see also Gozli et al., 2013) or in orthographically shallow languages, but it does occur under conditions that resemble those of Estes et al. (2008) – in English with short SOAs. The effect is also significant with or without a brief semantic context, though it tends to be larger with semantic context.

4. Further Consideration of Petrova et al. (2018)

Semantic Judgment of the Cue Words. Gozli et al.'s (2013) Experiments 4 and 6 both produced significant interference at short SOAs. Petrova et al. excluded these two studies from their meta-analysis on the basis that participants had to judge whether the cue (e.g., "2") belonged in a given category (e.g., "numbers"), and only on trials where the cue did belong in the category should they respond to the target stimulus. This cue judgment task violated Petrova et al.'s criterion that no task be performed on the cue words (see section 2 above). However, Petrova et al.'s Experiments 2 and 3 also required judgment of the cues (see Table S1). Those studies included "catch trials" in which the cue was a number word, and when participants saw a number word (e.g., "three"), they should first read aloud the cue word and then respond to the target stimulus. A reviewer of our manuscript argued that Petrova et al.'s task did not require judgment of the cue words. The reviewer argued that because participants needed to read aloud only the number words (catch trials), the other cue words (experimental trials) need not be judged. We believe it is impossible for participants to know that the experimental cue words were not number words (and therefore should not be read aloud) without judging whether they were number words. That is, to refrain from reading aloud a non-number word, one must first judge whether it is a number word. The reviewer further suggested that the proportion of catch trials (i.e., 6% in Petrova et al.'s experiments) is relevant to this issue, but we disagree. Even if there are only a few catch trials in the entire experiment, that would require a judgment of *all* cue words, or else participants would fail the catch trials. And as Petrova et al. report in their SOM, nearly all participants successfully completed the catch trials, indicating that nearly all participants successfully judged the cue words. Thus, if Petrova et al.'s Experiments 2 and 3 are to be included in the meta-analyses, then Gozli et al.'s Experiments 4 and 6 should also be included in the meta-analyses.

Simon Effect. Petrova et al. demonstrated that this spatial interference paradigm can additionally reveal a Simon effect, whereby the locations of the target stimulus and response key interact to affect responding. For example, because the "X" key is on a lower row of the keyboard than the "O" key, "X" and "O" responses are respectively faster to targets at the bottom or top of the visual display. Thus, any slight vertical displacement of the response keys, as in Estes et al.'s (2008) use of the "X" and "O" keys, can affect responding. However, spatial interference has been obtained several times with response keys that were not vertically displaced (Bergen et al, 2007, Experiments 1 and 2; Gozli et al, 2013, Experiments 3, 4, and 6; Petrova et al., 2018, Experiment 7), and several studies with vertically displaced response keys have failed to obtain spatial interference (Petrova et al., in press, Experiments 5 and 9; Renkewitz & Müller, 2015). Thus, vertical displacement of response keys is neither necessary nor sufficient for spatial interference, and hence the spatial interference effect cannot be explained as a Simon effect.

Experiment 10. In our main text we do not discuss Petrova et al.'s Experiment 10. In that experiment, participants read pairs of spatial words (e.g., "bird" → "cloud") and judged the spatial connotation of the second word ("indicate as fast and as accurately as possible whether the target word denotes a concept that usually appears at the top or at the bottom of the visual field"). Petrova et al. concluded from the results of this experiment that "The significant semantic priming effect rules out the possibility that cue word presentation conditions were inadequate for observing semantic effects."

In other words, Petrova et al. claim that the finding of semantic priming in their Experiment 10 provides evidence that semantic processing also occurred in their Experiments 1-9, yet the spatial interference effect did not occur. That conclusion, however, is not logically warranted. The task of Experiment 10, in which participants explicitly judge spatial associations, is clearly very different from all of Petrova et al.'s other experiments (and all experiments included in our meta-analysis), in which participants are not asked or required to judge spatial associations. Given the very different task used in this experiment, it is not informative of Petrova et al.'s other experiments. Showing that people are able to process words semantically – when instructed to – clearly does not indicate or imply that they do so when *not* instructed to process the words semantically. Finding semantic processing in a task that requires semantic processing simply cannot inform whether semantic processing occurs in some other, very different task that doesn't require semantic processing. Thus, we do not discuss Petrova et al.'s Experiment 10 because it is not informative of the spatial interference effect.

References

Bates, E., Burani, C., D'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, *29*(7), 986-999.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733-764.

Bond Jr, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*(4), 406-418.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333-342.

Estes, Z., Verges, M., & Adelman, J. S. (2015). Words, objects, and locations: Perceptual matching explains spatial interference and facilitation. *Journal of Memory and Language*, *84*, 167-189.

Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down: Object words orient attention to the objects' typical location. *Psychological Science*, *19*(2), 93-97.

Gozli, D. G., Chasteen, A. L., & Pratt, J. (2013). The cost and benefit of implicit spatial cues for visual attention. *Journal of Experimental Psychology: General*, *142*(4), 1028-1046.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099-1114.

Lebois, L. A. M., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science*, *39*, 1764–1801.

Ostarek, M., & Vigliocco, G. (2017). Reading sky and seeing a cloud: On the relevance of events for perceptual simulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 579-590.

Petrova, A., Navarette, E., Suitner, C., Sulpizio, S., Reynolds, M., Job, R., & Peressotti, F. (2018). Spatial congruency effects, just not for words: Looking into Estes, Verges, Barsalou (2008). *Psychological Science*.

Renkewitz, F. & Müller, S. (2015). Replication of Study 1 by Estes, Z., Verges, M., & Barsalou, L.W. (*Psychological Science*, 2008). Retrieved September 20, 2015 from https://osf.io/vwnit/

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin & Review*, *22*(6), 1614-1629.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559-569.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534-547.

Wilson, D. B. (2006). Meta-analysis macros for SAS, SPSS, and Stata. Retrieved July 21, 2017 from http://mason.gmu.edu/~dwilsonb/ma.html

Wilson-Mendenhall, C. D., Simmons, W. K., Martin, A., & Barsalou, L. W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience*, *25*(6), 920-935.