Dear Dr. Zoltan Dienes,

Thank you for allowing us to submit the revised version of our Stage-1 manuscript titled "Do task-irrelevant cross-modal statistical regularities induce distractor suppression in visual search?" to PCI RR.

We would like to thank you and the reviewers for their constructive comments and helpful suggestions. Below you can find our responses to all comments in bold.

We have submitted the revised Stage-1 Registered report (file name: "Registered_Report_Stage-1_Proposal_v2.pdf"). We have also uploaded a PDF document indicating modifications in Tracked changes.

We look forward to your and reviewer's comments.

With kind regards,
Kishore Kumar Jagini (on behalf of authors)

Recommender

I now have two reviews from experts about your submission. Both reviewers are overall positive, but they make a number of points that will need addressing in a revision. I want to draw your attention to three points in particular based on both my own reading and the reviewers' reactions, though all the reviewers' points need a response:

1) Align your statistical tests with the hypothoses tested. Vadillo asks about your ANOVAs. Note your Design Table does not refer to the ANOVAs, but to particular t-tests. Indeed, a valuable feature of the Registered Report format is you can plan precisely the contrast needed to test each hypothesis in advance. In order to limit inferential flexibility, other tests should typically not be specified. That is you do not need to specify omnibus ANOVA tests in order to justify the particular test of a hypothesis; one just specifies the exact contrast that tests each hypothesis. Further, in order to limit inferential flexibility, you should use just one system of inference: You could do frequentist t-tests or Bayesian ones; but pick one as the one you will do and from which inferences will follow.

Response: Thank you for the suggestions. We have aligned the planned statistical tests with the hypothesis in the revised version of manuscript. Further, to limit the inferential flexibility, we use only frequentist tests in the revised manuscript. Please see the updated text in the revised manuscript. Please see the uploaded PDF document indicating revision modifications in Tracked changes.

2) Power/sensitivity should be specified for each test with justification of the effect size chosen. Thus, if you use frequentist tests, you need to justify a minimally interesting effect size that is scientifically relevant for each test, then determine power for that test, indicating the power for each test. Vadillo asks where d = 0.6 comes from. See here for how to approach the problem of specifying an effect size for power. On the other hand you might decide to use Bayesian t-tests. Then you should justify the rough size of effect expected for each test; see previous reference for this too. The use of default scale factors especially for tests with few trials, like your awareness test, can lead to spurious support for H0 (see here).

Response: Thank you for the suggestions. In the revised manuscript we have provided the power for each test with justification of the effect size chosen. Please see the uploaded PDF document indicating revision modifications in Tracked changes.

3) Vadillo also questions the sensitivity of your test of awareness. This point is related to the previous ones. You need an appropriate sensitivity analysis of every test you conduct - and you also need to list your awareness test in the design table (remove description in the text of tests that you don't list in the design table, in order to keep inferential flexibility under control; you can always report these other tests in a non-preregistered results section in the final mansucript). See the "calibration" section of this paper for how to determine an expected effect size for an awareness test, or else the reference I gave at the end of point 2). The reviewer also brings up what the proper chance level is of your measure. Chance performance would be above zero.

Response: Thank you for the suggestions. In the revised manuscript, we have provided the power/sensitivity for awareness tests with justification of effect size chosen. We also removed the text of tests that are not listed in the design table.

<div align="center">

Reviewer #1

</div>

The registered report "Do task-irrelevant cross-modal statistical regularities induce distractor suppression in visual search? " is well written and has a valid research question. The logic and rationale of the proposed hypotheses are clear. Also, the methodology is sound, clear and replicable. The authors have also considered additional outcome-neutral conditions for manipulation checks. Some concerns are listed below.

On p.5, at the end of the first paragraph, the authors claimed that there seems to be enough evidence to support that our brain learns and utilize statistical regularities of both task-relevant and task-irrelevant sensory stimuli for optimizing behaviour. Given the authors only mainly introduced the influence of statistical regularities of the salient distractors in the previous part, they should make it more clear the evidence about task-relevant and task-irrelevant stimuli, respectively.

Response: We have provided the evidence about task-relevant statistical regularities in the first version (and also in the second version) of the manuscript (on p.4). The relevant text is mentioned below for the reference. We intended to keep minimal discussion about learning statistical regularities of the task-relevant stimuli as the current manuscript mainly focussed on learning statistical regularities of task-irrelevant stimuli.

"For instance, targets (task-relevant) that frequently appear at a particular spatial location in visual search displays are perceptually processed better than targets at infrequent search locations (Awh et al., 2012; Chun & Jiang, 1998; Geng & Behrmann, 2002, 2005; Jiang et al., 2013)."

The authors have included Chen et al.'s crossmodal contextual cueing studies which are quite relevant. However, I suggest the authors consult more crossmodal selective attention literature. For example, Spence C's lab has done a lot of studies on this topic.

Response: Thank you for the suggestion. However, we do not seem to find a literature from Spence C's Lab that is relevant to the "cross-modal statistical regularities and selective attention". So, we could not include the literature from the Spence C's Lab.

For the data analysis, the significance level alpha is set to 0.02. Why not use an alpha level of .05 that is commonly used to better balance the issues of Type I and Type II errors?

Response: We agree that it is a common practice to use the significance level (alpha/Type I error) as 0.05. However, some of the journals have stringent requirements regarding minimum required thresholds for pre-planned statistical evidence. For example, Cortex (one of the PCI RR friendly journals) requires alpha ≤ .02 and power ≥ 0.90 for all preregistered hypothesis tests to Stage 1 in-principle acceptance (IPA). Accordingly, we chose to set minimum thresholds: alpha ≤ .02 and power ≥ 0.90.

Overall, this is an interesting research proposal and I would like to see the outcome in the near future.

Response: Thank you.

The main goal of the two experiments proposed in this RR is to explore whether distractor inhibition in the additional singleton task can be modulated by contextual auditory information. Specifically, the singleton distractor will be present more frequently in two particular locations, each of then cued by a distinctive sound. The question is whether participants will learn to use this sound to suppress attention to the location whether the singleton distractor is most likely to appear in that trial. Although this type of contextual modulations have been explored in other visual statistical learning paradigms, this is the first time that such an effect is tested in the additional singleton task. I think it can be quite interesting for researchers working in this area. I have relatively minor comments about the general procedure, design and context that I think should be easy to tackle in a revised version.

The authors (and editor) won't be surprised to find out that I am a bit concerned about the awareness test included at the end of the experiment and the type of conclusions that can be drawn from them. The awareness tests included in this paradigm are almost always doomed to suggest that learning was unconscious. Participants' learning is assessed through hundreds of visual search trials using a continuous measure (reaction times), but their awareness is assessed briefly in 2-4 yes/no questions. Logically this procedure is much more sensitive to detect a significant effect in reaction times than to detect an equivalent effect in awareness, introducing a strong bias to conclude that learning was unconscious. To be completely honest, this is not the authors' fault: this kind of awareness tests is common in the area, but it happens to be highly misleading, and a well powered and carefully designed RR should avoid these shortcomings by all means. I am definitely not asking the authors to cite these papers, but to better appreciate these problems they might find useful to read our papers addressing this question in contextual cueing (Vadillo et al., 2016, PBR), location probability learning (Vadillo et al., 2020, JEP:Gen) and the additional singleton task (Vicente-Conesa et al., 2021, https://psyarxiv.com/yekvu/). In the particular case of location probability learning and the additional singleton task it is quite difficult to improve the sensitivity of the awareness test, because they are essentially a one-shot test, i.e., one cannot include more and more testing trials to improve the sensitivity of the awareness test, as can be done for instance in the contextual cueing task. But at least, one can try to complement the traditional yes/no dichotomous responses by continuous and potentially more sensitive measures. For instance, in the location probability learning task we have found that asking participants to rate the percentage of times the target has appeared in each quadrant is a more sensitive test than simply asking them to select a quadrant (e.g., Giménez-Fernández et al., 2020, JEPHPP). It is still a quite suboptimal measure, but slightly more sensitive. The authors can also consider replacing their yes/no or discrete choice responses by confidence ratings or any other response that provides a more nuanced and graded measure of awareness.

Response: Thank you for pointing out the methodological shortcomings in the manuscript. We have read through the suggested papers and convinced that the confidence rating scale and ranking methods are more sensitive measures for testing awareness. Accordingly, we have modified our methodology for awareness tests to include confidence rating scale and ranking method in the revised manuscript. Please see the uploaded PDF document indicating revision modifications in Tracked changes.

In any case, even if the authors decide to stick to this procedure (which I strongly advise against) I would still ask the authors to describe in much more detail what analyses they are planning to run on their awareness data. For the first and third questions (which are not very informative; this particular type of subjective rating is known to conflate "unawareness" with a conservative bias; see Flemming and Lau, 2014, https://www.frontiersin.org/articles/10.3389/fnhum.2014.00443/full), the authors plan to estimate the proportion of "yes" responses. But for the second and fourth question I don't think they provide sufficient information to understand how they are planning to process and analyze these responses. They say that they will calculate the "distance between the locations indicated by the participant and the actual locations". But how are they planning to do this? Recall that there are actually two high-probability locations and participants are selecting two locations. Let's imagine that the actual locations are, say 1 and 5, and the participant chooses 2 and 3, for instance. How is this reduced to a single distance score? In addition, the authors say that they will analyze these scores by comparing them against zero. But I can't understand the logic of this analysis. Shouldn't the authors compare the observed score against the score that would be ideally observed if responses were completely random? I think that the authors need to provide much more detail here.

Response: Thank you for pointing out the issues in the manuscript. We have modified our methodology for awareness tests to include confidence rating scale and ranking method to avoid shortcoming of methodological issues in the awareness tests. Kindly see the updated text for the awareness tests' questionnaire, analysis used in the revised manuscript. In the revised manuscript, we tried to address the concerns raised in the above comments.

Power. The sample size was calculated to provide reasonable power to detect a $d = 0.6$ effect. But why is this effect size a good reference? I am sure that distractor suppression in the additional singleton task is usually much larger than this, but do we have any evidence to expect that the contextual (auditory) modulation of the effect, if true, will be larger than this? In relation to my previous note, the study also plans to determine whether learning was unconscious. Knowing that responses to the awareness test are likely to be quite noisy (see my previous paragraphs) would $N = 39$ be enough to test this hypothesis with sufficient power? I honestly doubt it. Our meta-analysis of awareness in the contextual cueing task yielded an average effect of $dz = .31$, and there are good reasons to suspect that the typical awareness test in the contextual cuing task is much more sensitive than the traditional test in the additional singleton task (e.g., it usually includes around 24 trials instead of one-shot responses). For the probabilistic cuing task we found an average awareness effect of $h = .35$, which requires at least 64 participants to reach just 80% power.

Response: In the revised manuscript, we have provided the updated sample size with the justification of effect size for each tests. For your reference, we are copy pasting the relevant text below.

"Number of Participants:

For each proposed experiment in this study, we aim to recruit a minimum of 68 participants (who meets the participant selection criteria) from the Indian Institute of Technology. Given an effect size of d = 0.602 in a similar study by Failing et al. (2019), which was obtained by taking a difference between colour-match and colour-mismatch trials at two high probability distractor locations, a minimum of 39 participants required for power ≥ 90% with alpha set to 0.02 (calculated using G*Power 3.1). However, we do not have evidence to expect a similar effect in the context of auditory stimulus induced effect between two high probability distractor location trials. We believe that the effect size in the current experimental context is smaller than the effect size found in the study by Failing et al. (2019), and thus we are attempting to detect a smaller effect size of d = 0.45. The effect size d = 0.45 requires a minimum of 68 participants for each proposed experiment to get power ≥ 90% with alpha set to 0.02 (calculated using G*Power 3.1) in a two-tailed matched-sample t-test. This sample size is considerably larger than the typical experiments conducted using the additional singleton tasks (an average of around 26 participants in (Failing et al., 2019; Wang & Theeuwes, 2018a, 2018b, 2018c)).

Justification for the sample size to determine awareness of statical regularities in the proposed experiments: A minimum of 68 participants for each proposed experiment will be considered for testing awareness about the relationship between auditory and visual distractor location statistical regularities. Most previous studies utilized dichotomous "Yes" or "No" responses and/or indicating a particular location where participants believe that the target/distractor appeared most frequently to test awareness about statistical regularities and concluded that the statistical learning is unconscious (e.g., in studies by (Failing et al., 2019; Wang & Theeuwes, 2018b)). However, recent studies indicated that using a confidence rating scale and ranking methods are, arguably, more sensitive measures for testing awareness (Giménez-Fernández et al., 2020; Vadillo et al., 2020). Utilizing these sensitive measures to test awareness of statistical regularities in probabilistic cuing search tasks, the Vadillo et al. (2020) study indicated that participants are not unaware of the statistical regularities. Their study reported an effect size of Cohen's h = 0.52 in their experiment 1. However, we do not have evidence to expect a similar effect size, utilising these sensitive measures of awareness, in studies identical to the proposed experimental context. Assuming that the effect size in the current experimental context might be small, we are attempting to detect a smaller effect size of d = 0.45. The effect size of d = 0.45 requires a minimum of 68 participants for each proposed experiment to get power ≥ 90% with alpha set to 0.02 (calculated using G*Power 3.1) in a two-tailed matched-sample t-test."

Minor comments

"task-irrelevant" -> the auditory stimuli are characterized in the ms as "task-irrelevant". But given that they actually convey useful information (i.e., where the distractor is going to appear) I wonder if the name is actually fair. Wouldn't it be better to refer to these stimuli as "contextual" stimuli instead?

Response: We have defined the task-relevant stimuli mainly for stimuli that matches the explicit task goals (i.e., search target), and any other stimuli regarded as "task-irrelevant". Therefore, we would prefer the auditory stimuli are characterized as "task-irrelevant auditory stimuli" or "task-irrelevant auditory context"

At several points, the ms gives the impression that the studies explore how participants "anticipate" the distractors, i.e., how they are "perceptually suppressed by pro-active modulations" (p. 6). If the goal is to study anticipatory behavior, shouldn't the auditory signals be presented before the search display instead of simultaneously?

Response: In the manuscript we have mentioned that the auditory stimulus is presented simultaneously with the visual display. However, we need to account for the neural/perceptual processing differences between auditory and visual signals. For example, processing of simple auditory pip can be approximately 50ms faster than a visual flash (Vroomen & Keetels, 2010). Moreover, in the proposed experiments, there are multiple visual stimuli in the search display which can further delay the processing time compared to a single flash. In such a scenario, we assume that auditory signal might sufficiently processed earlier in time to provide an anticipation regarding distractor locations in the context of proposed experimental design. In the cross-modal attention literature, studies have also suggested that the visual search for target benefited when the search displays are presented simultaneously with the auditory signal (e.g., Burg et al., 2008; Iordanescu et al., 2010).

p. 7 "HpValD" and "HpInVald". The meaning of these acronyms only becomes clear in the following pages. Wouldn't it be easier to understand the Hypotheses section if the previous paragraphs introduced the designed briefly, including the condition names?

Response: Thank you for the suggestion. Please see the uploaded PDF document indicating revision modifications in Tracked changes.

p. 7 I found it a bit weird that the authors present Hypothesis 1:2 as an additional hypothesis. It is simply the negation of Hypothesis 1:1, isn't it? Same comment in the Study Design Table.

Response: We are sorry for the confusion. We have now removed the Hypothesis 1:2 section in the updated manuscript.

p. 7 "… the former condition associated with the search trials… should produce fater RTs" Faster compared to what? Same problem in the following sentence.

Response: We have corrected the sentences in the revied manuscript to avoid the confusion. Thank you for pointing out.

p. 13. Note that the singleton distractors appear much more frequently in two locations than in any of the other locations. This is unavoidable, of course, but it is important to remember that it renders some of the statistical comparisons meaningless. For instance, there is no reason to compare either HpValD or HpInvalD with the low probability distractor location. Any significant difference could be due to either to the fact that the sound is unpredictive in the latter condition or to the fact that the distractor has appeared in a location where it seldom appears. It will be quite difficult to interpret this result. So, I wonder if it makes sense to include all four conditions in a single ANOVA (p. 17)

Response: Thank you for the suggestion. We have removed the omnibus ANOVA from the planned tests. We will use paired sample t-test to compare HpValD or HpInValD conditions.

p. 19. In the awareness test, will participants understand what "colored non-target locations" are?

Response: We have modified the questionnaires for the awareness tests. Please see the uploaded PDF document indicating revision modifications in Tracked changes.

Appendix:
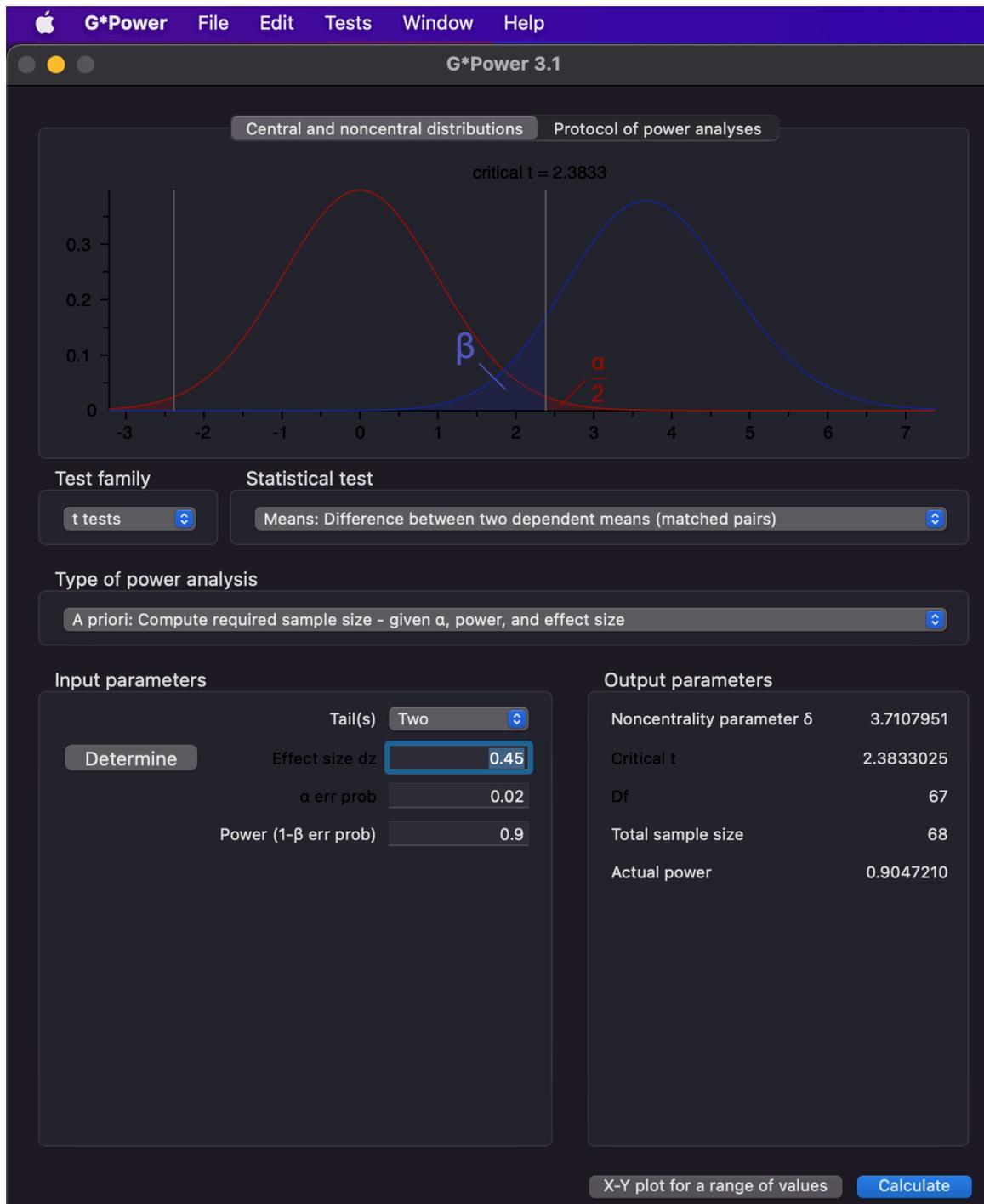
References:

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Burg, E. V. der, Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

Chun, M. M., & Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Failing, M., Feldmann-Wüstefeld, T., Wang, B., Olivers, C., & Theeuwes, J. (2019). Statistical Regularities Induce Spatial as well as Feature-Specific Suppression. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(10), 1291–1303. https://doi.org/10.1037/xhp0000660

Geng, J. J., & Behrmann, M. (2002). Probability Cuing of Target Location Facilitates Visual Search Implicitly in Normal Participants and Patients with Hemispatial Neglect. *Psychological Science*, *13*(6), 520–525. https://doi.org/10.1111/1467-9280.00491

Geng, J. J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics*, *67*(7), 1252–1268. https://doi.org/10.3758/bf03193557

Giménez-Fernández, T., Luque, D., Shanks, D. R., & Vadillo, M. A. (2020). Probabilistic cuing of visual search: Neither implicit nor inflexible. *Journal of Experimental Psychology. Human Perception and Performance*, *46*(10), 1222–1234. https://doi.org/10.1037/xhp0000852

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*(7), 1736–1741. https://doi.org/10.3758/app.72.7.1736

Jiang, Y. V., Swallow, K. M., Rosenbaum, G. M., & Herzig, C. (2013). Rapid Acquisition but Slow Extinction of an Attentional Bias in Space. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 87–99. https://doi.org/10.1037/a0027611

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, *149*(1), 160–181. https://doi.org/10.1037/xge0000632

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics*, *72*(4), 871–884. https://doi.org/10.3758/app.72.4.871

Wang, B., & Theeuwes, J. (2018a). How to inhibit a distractor location? Statistical learning versus active, top-down suppression. *Attention, Perception, & Psychophysics*, *80*(4), 860–870. https://doi.org/10.3758/s13414-018-1493-z

Wang, B., & Theeuwes, J. (2018b). Statistical Regularities Modulate Attentional Capture. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(1), 13–17. https://doi.org/10.1037/xhp0000472

Wang, B., & Theeuwes, J. (2018c). Statistical regularities modulate attentional capture independent of search strategy. *Attention, Perception, & Psychophysics*, *80*(7), 1763–1774. https://doi.org/10.3758/s13414-018-1562-3