Critique of Leutgeb et al. studies

In two studies, Leutgeb et al. claimed that IVET *increases* the LPP at a late interval to spiders in adults (800–1500 ms, Leutgeb et al., 2009) and children (600–1200 ms, Leutgeb et al., 2012) during passive picture viewing. In addition to our main concerns (see manuscript), we identified several additional concerns:

If the gaze patterns change as a result of therapy, any changes in LPP are confounded by differences in gaze patterns before therapy compared to after therapy. Critically, this would mean that the actual processes that are indexed by LPP might be completely unaffected by therapy and that any changes in LPP are only an indirect effect of changes in gaze. So, if the claim is correct (Leutgeb et al., 2012; Leutgeb et al., 2009), then no conclusions can be drawn about any actual processes indexed by LPP because neither study controlled gaze during the task. Importantly, none of the studies (Leutgeb et al., 2012; Leutgeb et al., 2009) measured gaze directly and thus, there is no direct evidence to support the claim.

Furthermore, Leutgeb et al. (2009) admitted that the apparent increase in LPP in adults after therapy was opposite to their original hypothesis. As such, the data are only hypothesis generating (exploratory) and not hypothesis testing (confirmatory) (Nosek et al., 2018; Wagenmakers et al., 2018). Because the given explanation is post hoc, it is only tentative until supported by independent findings (Nosek et al., 2018).

Another concern is that several analytic decisions suggest that the reported findings (Leutgeb et al., 2012; Leutgeb et al., 2009) are not as robust as desired. With regard to the study by Leutgeb et al. (2009), the authors calculated topographical difference maps for the contrasts of interest and "selected electrode sites for statistical analyses based on these maps" (p. 294). Thus, the data were used twice: To decide about which electrodes and intervals should be analyzed for an effect, and to test this effect statistically. However, this practice of *double dipping* results in a nonindependence error and inflates the risk for false positives (Kriegeskorte et al., 2009; Makin & Orban de Xivry, 2019). Relatedly, LPP was arbitrarily divided into early and late LPP, and each interval was tested in a separate ANOVA. Thus, these multiple comparisons may have increased the risk for false positives (Luck & Gaspelin, 2017; Makin & Orban de Xivry, 2019). Furthermore, the LPP was analyzed only for a single electrode, and the selected electrode differed between analyses: When comparing patients with controls before treatment, only Pz was analyzed, whereas when comparing patients before and after treatment, only Cz was analyzed. Because the LPP is characterized by a wide-spread central-parietal positivity (Hajcak et al., 2011; Wiens et al., 2012), the neighboring electrodes Cz and Pz are relatively close and should be expected to show similar sensitivity to any effects. Thus, there is no a priori reason for this electrode switch. Taken together, these analytic decisions likely increased the risk for false positives (Gelman & Loken, 2013; Hoffmann et al., 2021).

Also, consistent with the claimed treatment effect on late LPP, Figure 2D in Leutgeb et al. (2009) suggests that the only noteworthy effect was that amplitudes increased to spiders for the treatment group. However, this figure does not include amplitudes to neutral pictures, but these are important as a baseline (as LPP is a difference, spider–neutral). Notably, Figure 2B in Leutgeb et al. (2009), which includes the amplitudes to neutral pictures, suggests a baseline shift in that mean amplitudes to neutral pictures increased from session 1 to session 2 in the waitlist group. Although this effect was not significant ($p > .30$), it contributed to the three-way interaction between group (treatment group, waitlist), time (session 1, session 2), and category (spider, neutral). Specifically, an apparent increase of the amplitude difference between spiders and neutral pictures from session 1 to session 2 in the treatment group was strengthened by a concurrent decrease of the amplitude difference between spiders and neutral pictures from session 1 to session 2 in the waitlist group.

We are also concerned with regard to the study by Leutgeb et al. (2012) that examined effects of therapy in spider-phobic girls. Although we briefly describe our main concern in the manuscript, we elaborate it here for clarification. We argue that if anything, the effect was opposite to what

the authors claimed: LPP decreased rather than increased after therapy. In the study, Leutgeb et al. focused on mean amplitudes at Fz. Figure 3 in Leutgeb et al. (2012) suggests that the mean Fz amplitudes to spiders increased from session 1 to session 2 only for the treatment group. In support, an ANOVA showed a significant three-way interaction of the mean Fz amplitudes between group (treatment group, waitlist), (session 1, session 2), and category (spider, neutral). Also, follow-up t tests showed that the session effect (1 vs 2) was significant only for spiders in the treatment group. At face value, these results seem consistent with the claim that mean amplitudes increased to spiders only after therapy.

However, the main problem is that the amplitude difference of spiders minus neutral pictures at Fz is negative rather than positive. As such, this *negative* LPP (of spiders – neutral) does not fit the definition of the LPP as a late *positive* potential. Because Leutgeb et al. (2012) used an average reference, the topography of the LPP should be positive for central-parietal electrodes (i.e., Cz and Pz) but negative for other electrodes (e.g., Fz), as the mean of all electrodes is set to zero (Hajcak et al., 2011; Wiens et al., 2012). In support, in a similar study with spider-phobic girls, an average reference was used, and LPP was apparent as a clear positivity at Pz and a negativity at Fz (Leutgeb et al., 2010). Figure 1 in Leutgeb et al. (2012) shows exactly this pattern for the difference between spiders and neutral pictures in session 1: a relative positivity at Pz and a relative negativity at Fz. Therefore, we argue that the reported negativity at Fz reflects only the polarity reversal of the LPP and that the primary measure of the LPP ought to be the amplitudes at Pz.

From this perspective, the finding that amplitudes at Fz *increased* (i.e., became less negative) from before to after therapy means that (after correcting for the polarity reversal) LPP *decreased* from before to after therapy. Figure 1 in Leutgeb et al. (2012) suggests this pattern in that amplitudes at Pz were lower after therapy, but the three-way interaction at Pz was not significant.

In sum, whereas the main electrode for the LPP (i.e., Pz) does not suggest any changes in LPP, the pattern of results at the electrode that picks up the polarity reversal of the LPP (i.e., Fz) suggests that LPP decreased (rather than increased) with treatment. To conclude, the results by Leutgeb et al. (2012) suggest that if anything, LPP to spiders decreased rather than increased. This conclusion is opposite to that of the authors and opposite to the apparent results of their earlier study (Leutgeb et al., 2009).

Additional methodological concerns about the study by Leutgeb et al. (2012) are that although the design was similar to that of earlier studies (Leutgeb et al., 2010; Leutgeb et al., 2009), electrodes and intervals were defined differently. Also, electrodes (Fz, Cz, and Pz) were analyzed separately rather than as an additional independent variable in the ANOVA. These decisions may have increased risks for false positives (Gelman & Loken, 2013; Hoffmann et al., 2021; Kriegeskorte et al., 2009; Luck & Gaspelin, 2017). Furthermore, results for mean amplitudes at Fz suggested that in session 1, spider pictures did not differ from disgust pictures in the treatment group, and spider pictures did not differ from fear and disgust pictures in the waitlist group. At face value, this lack of differences between spiders and other negative but less arousing pictures suggests that there was no clear manipulation check; that is, responses to spiders may have differed only from neutral but not from other negative but less arousing pictures. Last, analyses of the P300 yielded significant three-way interactions at Fz and Cz and also significant follow-up t tests (e.g., the waitlist group showed more negative P300 to spiders at Fz in session 1 than session 2). Nonetheless, the authors concluded that "there were no therapy-related changes in response to spider pictures in earlier time frames of the ERP (i.e., the P300)" (p. 103). But, because these results were obtained by the same analysis strategy as was used for the LPP, an unbiased approach would be to consider results for P300 and LPP similarly as either true positives or false positives.

In sum, the hypothesis by Leutgeb et al. (2009) that changes in LPP are driven by changes in gaze was post hoc. Also, it implies that any LPP results would be confounded unless an effort is made to control gaze. That is, the actual processes that are indexed by LPP might be completely unaffected, as any changes in LPP could simply be an indirect effect of changes in gaze. Importantly,

none of the studies (Leutgeb et al., 2012; Leutgeb et al., 2009) measured gaze directly and thus, there is no direct evidence to support the hypothesis. Also, the alleged increase in LPP at Fz by Leutgeb et al. (2012) captures the polarity reversal of the LPP, which peaks at Pz. Thus, the findings suggest that treatment decreases rather than increases the LPP. Furthermore, because of several methodological concerns, the findings (Leutgeb et al., 2012; Leutgeb et al., 2009) cannot be considered robust, unbiased evidence (Gelman & Loken, 2013; Hoffmann et al., 2021).

To conclude, the reported results (Leutgeb et al., 2012; Leutgeb et al., 2009) cannot be taken as support for the claim that treatment increases the LPP to spiders after therapy. In fact, if results are taken seriously, one study suggests that LPP increases (Leutgeb et al., 2009) whereas the other suggests that LPP decreases after therapy (Leutgeb et al., 2012). Thus, these studies do not resolve whether LPP changes after therapy.

References

Gelman, A., & Loken, E. (2013). The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis Was Posited Ahead of Time. *Department of Statistics, Columbia University.* http://www.stat.columbia.edu/_gelman/research/unpublished/p_hacking.pdf

Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2011). *ERPs and the Study of Emotion.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0222

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), rsos.201925, 201925. https://doi.org/10.1098/rsos.201925

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Leutgeb, V., Schäfer, A., Köchel, A., Scharmüller, W., & Schienle, A. (2010). Psychophysiology of spider phobia in 8- to 12-year-old girls. *Biological Psychology*, *85*(3), 424–431. https://doi.org/10.1016/j.biopsycho.2010.09.004

Leutgeb, V., Schäfer, A., Köchel, A., & Schienle, A. (2012). Exposure therapy leads to enhanced late frontal positivity in 8- to 13-year-old spider phobic girls. *Biological Psychology*, *90*(1), 97–104. https://doi.org/10.1016/j.biopsycho.2012.02.008

Leutgeb, V., Schäfer, A., & Schienle, A. (2009). An event-related potential study on exposure therapy for patients suffering from spider phobia. *Biological Psychology*, *82*(3), 293–300. https://doi.org/10.1016/j.biopsycho.2009.09.003

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Makin, T. R., & Orban de Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, *8*, e48175. https://doi.org/10.7554/eLife.48175

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13*(4), 418–427. https://doi.org/10.1177/1745691618771357

Wiens, S., Molapour, T., Overfeld, J., & Sand, A. (2012). High negative valence does not protect emotional event-related potentials from spatial inattention and perceptual load. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 151–160. https://doi.org/10.3758/s13415-011-0072-8