PSICOLÓGICA

Worse than expected: A z-curve reanalysis of motor cortex stimulation studies of embodied language comprehension

Pablo Solana[1,2] **and Julio Santiago**[1,2]

[1]*Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain.*
[2]*Department of Experimental Psychology, University of Granada, Granada, Spain.*

Abstract

Dozens of TMS and tDCS studies suggest a functional involvement of motor cortex in action language comprehension, supporting the embodied cognition view. In a recent study (Solana & Santiago, 2022, *Neurosci. Biobehav. Rev.*, *141*, 104834), we evaluated the soundness of this literature by means of *p*-curve analyses and tests for excess significance. The analysis estimated a low average power (≈ 30%) and showed signs of publication bias, which led us to conclude that this body of findings does not stand on solid ground. Yet, the employed techniques seem to perform poorly when high heterogeneity is present (as is the case) and cannot quantify the amount of publication bias. For these reasons, in the present paper, we reanalyzed the same set of studies with a method that does not have these limitations: z-curve analysis. Z-curve not only replicated our prior conclusions but showed an even more heartbreaking situation, with a lower power estimation (≈ 20%) and clear signs of a strong publication bias (around 6-10 unpublished contrasts expected to exist for each published one). Researchers on this topic should consider these shortcomings before drawing conclusions from this body of findings, as well as start implementing robust and transparent research practices to improve their reliability.

Keywords: Embodied cognition, Language comprehension, Motor system, TMS, tDCS, Z-curve analysis, Reanalysis.

Introduction

Which neurocognitive mechanisms underpin action language comprehension? According to the grounded and embodied theories, conceptual processing is closely linked to our bodily experiences (for a review, see Barsalou, 2008). On this basis, several influential versions of embodiment assert that the understanding of action-related language requires simulating its motor content via the motor areas of the brain (e.g., Barsalou, 1999; Pulvermüller, 2005). This view thereby predicts that motor activation is an automatic and necessary component of action language semantics. Consider the verbs *pick* and *kick*. The mentioned account would predict that the hand portion of the motor cortex will be recruited to process *pick*, whereas the foot motor cortex will be engaged in processing *kick*. A number of behavioral (e.g., Glenberg et al., 2008), EEG (e.g., Hauk & Pulvermüller, 2004), and neuropsychological (e.g., Boulenger et al., 2008) studies have concluded in support for this idea. Despite that, whether the motor system is causally involved in language understanding remains under debate (e.g., see Mahon & Caramazza, 2008).

Psicológica (2023)

DOI
10.20350/digitalCSIC/15661

Corresponding author
Pablo Solana
solana@ugr.es

Edited by
Miguel A. Vadillo

Reviewed by
Alejandro Sandoval-Lentisco

© Solana and Santiago, 2023

A key strategy to test the causal involvement of motor areas in language comprehension has been the use of non-invasive brain stimulation techniques such as Transcranial Magnetic Stimulation (TMS; e.g., Buccino et al., 2005; Repetto et al., 2013; Vukovic et al., 2017) and Transcranial Direct Current Stimulation (tDCS; e.g., Birba et al., 2020; Vitale et al., 2021). On the one hand, TMS to the motor cortex elicits motor-evoked potentials (MEPs) that can be recorded from the muscles. This has been used to assess whether the motor system is recruited within the temporal window at which semantic processing is expected to occur. For instance, Buccino and colleagues (2005) found that listening to hand- and foot-related sentences decreased the amplitude of MEPs from hand and foot muscles, respectively, 200 ms after sentence presentation. On the other hand, repetitive TMS protocols (rTMS) and tDCS allow to disrupt the normal functioning of motor cortex (similar to a brain lesion) and test how this impacts language comprehension. If the cortical motor areas play a causal role in linguistic processing, then stimulating those regions ought to affect how people process language. As an example, studies like Repetto et al. (2013) or Vukovic et al. (2017) have found that the application of rTMS to the hand primary motor cortex slowed down responses for manual (vs. non-manual) verbs in semantic judgement tasks. This kind of brain stimulation studies have been considered as clear causal evidence in support of the embodiment hypothesis (e.g., see Barsalou, 2008).

In a recent paper (Solana & Santiago, 2022; SS22 hereafter), we assessed the evidential value, average power, and publication bias of this literature. To do so, we first searched for all the published studies that used non-invasive brain stimulation (TMS and tDCS) to test the implication of motor cortex in action language processing. Forty-three studies were selected. Then, we submitted them to two meta-analytic techniques: *p*-curve analysis (Simonsohn et al., 2014) and test for excess significance (TES; Ioannidis & Trikalinos, 2007). *P*-curve is a method that relies on how the significant *p*-values of a set of studies distribute: If there are no true effects in a given literature, the distribution should be flat. If the majority of the effects are real, then the distribution should exhibit more values close to $p = 0$ than to $p = 0.05$. And under extreme forms of *p*-hacking, values close to 0.05 can even be more frequent than those close to 0. *P*-curve is thus able to inform about both the evidential value of a set of studies (i.e., to what extent the findings can be explained by the presence of non-optimal practices) and their underlying average statistical power (i.e., their expected replication rate if identically repeated). TES is used to compare the observed proportion of significant findings in the study set with its average power (in our study, identified by the *p*-curve). A significantly greater number of observed than expected significant findings is interpreted as a sign of publication bias.

The results were striking. First, the shape of the *p*-curve showed that, after two decades of research and over 40 studies, the majority of them reporting significant findings, the evidential value of this set of studies is currently inconclusive. Second, their estimated underlying power was quite low (≈ 30%),

which predicts that around 70% of them would not replicate if repeated identically. Third, TES suggested publication bias, as indexed by a disproportionate number of significant results (≈ 70%) given their estimated power (≈ 30%). In a nutshell, our analyses indicated that the published motor cortex stimulation studies of embodied language comprehension do not stand on solid ground.

Yet, *p*-curve analysis and TES are not exempt of criticisms. Although *p*-curve analysis was designed to deal with heterogeneity (Simonsohn et al., 2014), some authors have shown that it overestimates power under conditions of high heterogeneity (McShane et al., 2016; Brunner & Schimmack, 2020; van Aert et al., 2016; but see Simmons et al., 2018). This could have influenced our results, as there are large differences in the experimental procedures and measures within these studies (see SS22). If so, one can suspect that the situation is even worse than concluded in SS22. In a similar vein, TES has also been shown to produce biased outcomes in presence of heterogeneity (Renkewitz & Keiner, 2019). Moreover, since we used the power estimated by the *p*-curve to run the TES, the limitations of the former technique may also have contributed to bias the conclusions from the latter. Finally, it is also worth mentioning that TES can reveal the presence of publication bias but cannot quantify the amount of bias (Bartoš & Schimmack, 2022).

Fortunately, a recently developed tool seems to be able to circumvent these limitations: *z-curve analysis* (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). Broadly, z-curve takes both significant and non-significant *p*-values as input (unlike *p*-curve, which only evaluates significant values), recomputes them into z-scores, and models them as a mixture of truncated folded normal distributions. As a result, not only does it yield better estimations than *p*-curve and TES when heterogeneity is present, but it also provides more revealing plots and allows quantifying the amount of publication bias (see the Methods section for details).

In the present paper, we reanalyze the studies included in SS22 by means of z-curve analysis. This provides a less biased and more complete evaluation of the reliability of the extant TMS and tDCS studies assessing the involvement of motor cortex in action language comprehension.

Methods

Transparency and open practices

Raw data, analysis scripts, high quality plots, and supplementary information of the present study are available at the following Open Science Framework (OSF) repository: https://osf.io/4f6em/

Contrast selection

We reanalyzed the same set of studies included in SS22, which includes 43 studies containing 47 experiments (e.g., Birba et al., 2020; Buccino et al., 2005; Repetto et al., 2013; Vitale et al., 2021; Vukovic et al., 2017). All of them were peer-reviewed studies published in international journals that used either TMS or tDCS over cortical motor areas (primary motor cortex, premotor cortex, or supplementary motor area) and employed tasks related to the processing of action-related words (verbs, nouns, and adjectives), sentences, or texts. The dependent variables of the studies included behavioral measures such as reaction time, accuracy, recall rate, or learning measures, but also neurophysiological measures like MEPs. The complete list of selected studies, as well as their main characteristics (e.g., type of task, dependent variable, or sample size), can be consulted in Table 1 of SS22.

The contrasts selected were the same as in SS22, with the exception that we included several non-significant results that could not be included in our prior *p*-curve analysis. All the contrasts selected refer to the central hypotheses of the analyzed studies and were selected following the guidelines provided by Simonsohn et al. (2014). The exact *p*-value of all the newly included contrasts was recalculated from the reported statistic through the *P*-Curve App 4.06 (http://www.p-curve.com/app4/). None differed from the value reported in its respective paper. Yet, some non-significant results were reported in a vague way without declaring their exact numerical value (e.g., "No other effects reached significance") or without the value of the statistical test they refer to (e.g., *F* or *t*). This impeded us recalculating their exact value and therefore, inputting them into the z-curve analysis. In these cases, we emailed the corresponding authors of the papers and asked them for the missing information. Only 2 out of 8 authors replied, and neither of them reported having access to the requested data anymore.

On many occasions, more than one valid contrast could be selected for a study (e.g., in studies with more than one hypothesis, or studies testing a hypothesis with several tests). In these cases, to avoid selection biases, and as Simonsohn et al. (2014) recommend, SS22 ran two complementary analyses using two slightly different sets of contrasts: a main analysis and a robustness analysis. Here, we adopted the same strategy. When two or more hypotheses were present in a paper, the contrast testing the first hypothesis appearing in the paper was selected for the main analysis, and the contrast testing the second hypothesis was included in the robustness analysis. The same applies to those papers testing the same hypothesis in several ways: the first test reported was selected for the main analysis, while the second one was included in the robustness analysis.

In brief, 53 contrasts were initially selected for both the main and the robustness analysis, but 15 (4 presumably significant and 11 presumably non-significant) were excluded from the main analysis, and 13 (4 presumably significant and 9 presumably non-significant) from the robustness analysis, because their exact value was not reported or could not be recalculated. As a result, 38 values (33 significant and 5 not significant; coming from 33 studies) were included in the main analysis, and 40 values (33 significant and 7 not significant; coming from 33 studies) were included in the robustness analysis. An updated version of SS22's disclosure table including all the values entered in this reanalysis can be found in the OSF repository (https://osf.io/7w8c4).

Analytic strategy: Z-curve analysis

Z-curve analysis transforms a set of *p*-values (including both significant and non-significant values) into z-scores. When there are no true effects in a set of studies, z-scores distribute following a normal distribution centered in z = 0 and the probability of obtaining a significant result is 5% (assuming an alpha level of 0.05). However, when true effects underlie a literature, the likelihood of obtaining a significant result depends upon the statistical power we have: the more power, the more likely it is to obtain a significant *p*-value (Lakens, 2022a). Based on these assumptions, z-curve takes a set of z-transformed *p*-values and models them as a mixture of truncated folded normal distributions to calculate several indexes of interest (for details, see Bartoš & Schimmack, 2022).

First, it provides two indexes of statistical power: the Expected Discovery Rate (EDR) and the Expected Replicability Rate (ERR). The EDR is the expected proportion of significant results in the set of studies, including both studies reporting significant and non-significant findings. In other words, the estimated average power of all the studies included in the analysis. The ERR is the expected proportion of significant results only within the studies reporting significant findings. That is, the estimated power and replicability rate of the studies with significant findings.

Second, z-curve provides indexes of publication bias. To do so, the EDR can be compared with the Observed Discovery Rate (ODR): the observed proportion of significant results. If the estimate of the ODR does not lie within the 95% CI of the EDR, then the analysis suggests the presence of publication bias. Moreover, and contrary to TES, z-curve analysis also allows to quantify the amount of publication bias. Dividing the ODR between the EDR reveals the File-Drawer Rate: the estimated number of unpublished results for every published one.

The reanalysis was performed in R (R Core Team, 2021) through the zcurve package (version 2.3.0; Bartoš & Schimmack, 2020). The raw data and the analysis script are available in the OSF repository.

Results

What is the power underlying these studies? The EDR estimated from the model was 0.14 (95% CI [0.05, 0.31]) for the main analysis and 0.09 (95% CI [0.05, 0.31]) for the robustness analysis, which indicates an average power of around 10% for all the conducted studies, including those with significant and non-significant results. The ERR was estimated at 0.18 (95% CI [0.03, 0.39]) for the main analysis and 0.20 (95% CI [0.03, 0.43]) for the robustness analysis. This suggests an average power of around 20% for the studies reporting significant findings. In other words, if we conducted exact, direct replications of the studies reporting significant results, only around 20% of these studies would replicate.

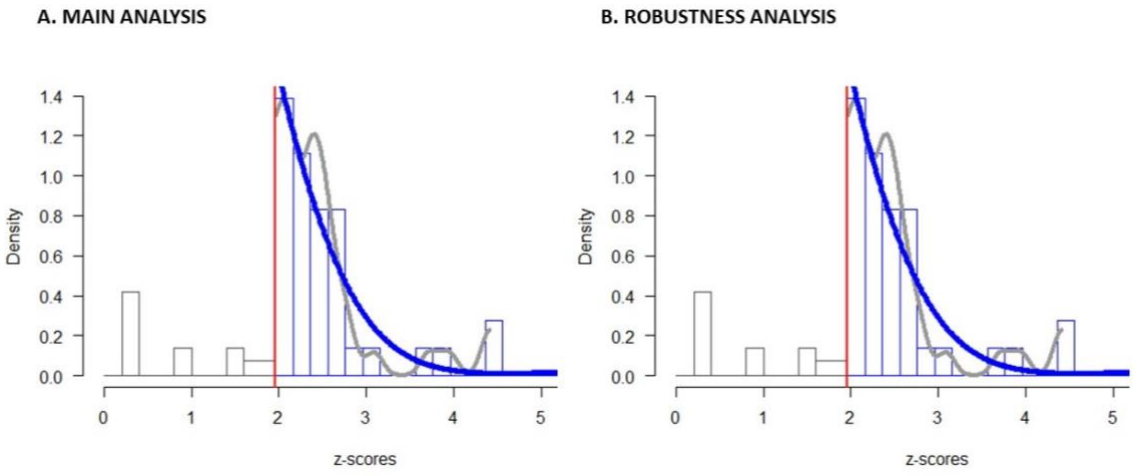

Figure 1. *Z-curve plot of the articles included in Solana and Santiago (2022). The histograms represent the distributions of the z-scores. The red vertical lines represent a z-score of 1.96, which corresponds to the significance threshold of p = 0.05 (two tailed). The blue continuous lines represent the estimated distributions of the inputted values. The blue dotted lines represent the 95% CI for the expected distributions. The plot to the left refers to the main analysis, whereas the plot to the right refers to the robustness analysis (see the OSF repository for a higher quality version of the plot).*

Is there publication bias within these studies? Figure 1 depicts the observed (histograms) and expected (blue lines) distribution of *p*-values (transformed to z-scores) for the main and the robustness analysis. In both cases, the plot clearly shows a much lower number of observed than expected non-significant values (those to the left of the red line representing the significance threshold). This indicates the existence of several missing contrasts, suggesting the clear existence of publication bias. The latter conclusion is also supported by the comparison between the EDR and the ODR. The ODR was 0.87 (95% CI [0.71, 0.95]) for the main analysis and 0.82 (95% CI [0.67, 0.92]) for the robustness analysis. Crucially,

the estimates of the ODRs are far from being captured by the 95% CIs of the EDRs, reaffirming the presence of publication bias. Moreover, the File-Drawer Rate was estimated at 6.38 (95% CI [2.21, 19.00]) for the main analysis and 10.13 (95% CI [2.23, 19.00]) for the robustness analysis, which predicts that there are approximately between 6 and 10 unreported non-significant contrasts for each reported significant contrast.

Yet, our results might be somewhat biased by the fact that the above-presented ODRs (and consequently the File-Drawer Rates) only include those *p*-values that could be recalculated. To check for this possibility, we computed an additional ODR by manually counting the number of significant vs. non-significant contrasts, independently of whether we could recalculate them or not (as we also did in SS22). The ODR value was 69.8% for both the main and the robustness analysis. As with the ODRs calculated through the z-curve analysis, this value is not captured by the 95% CI of the EDRs. The resulting File-Drawer Rate was 4.99 for the main analysis and 7.76 for the robustness analysis. Therefore, these results yield the same conclusions than using the z-curve ODR: that this set of studies contains signs of publication bias.

Discussion

In this paper, by means of z-curve analyses, we reanalyzed the data from a recent study (Solana & Santiago, 2022; SS22) in which we challenged the reliability of the available motor cortex stimulation studies of embodied language comprehension using *p*-curve analyses and tests for excess significance (TES). Broadly, the present z-curve reanalysis supports our previous conclusions by showing that, indeed, this set of studies has low statistical power and contains signs of publication bias. Actually, it suggests that these problems are even worse than reported in SS22.

First, in SS22, we estimated an average statistical power for this literature at around 30%. Here, z-curve estimated power at around 20% for the studies reporting significant findings, and only around 10% for all the extant studies (those reporting both significant and non-significant findings). Therefore, if we conducted identical direct replications of these stimulation studies, the vast majority of them would not replicate (≈ 80%), posing clear problems for embodiment research. Importantly, since z-curve is more robust to heterogeneity than *p*-curve (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020), this power estimation is less likely to be biased than the one obtained in our previous *p*-curve. As discussed in detail in SS22, we think that the cause beneath this low power is probably the use of small sample sizes, which has been identified as one of the main reasons of low-powered studies in cognitive science (e.g., Button et al., 2013). In this regard, we already noticed that the mean sample size of these studies is just N = 22.7 and that, with very few exceptions, these sample sizes are not based on a-priori power analyses (see SS22).

Second, the TES in SS22 showed a clear disparity between the observed and expected rates of significant results, suggesting the presence of publication bias. Z-curve not only replicates the significant disparity found in our prior TES but also estimates that there are approximately between 6 and 10 unreported non-significant results for every reported significant result (or between 5 and 8 if considering the ODRs that include those values that cannot be recalculated). This shocking index, together with the visual inspection of Figure 1, clearly suggests that there is a large amount of publication bias in this literature.

Still, the causes of this bias remain unclear, and present analyses cannot discern between the possible causes. One possibility is that only those studies with positive findings have been published (Rosenthal, 1979), which predicts that there may be more than 300 unpublished studies on this topic (around 6-10 per every published one). Considering the difficulties and costs of conducting brain stimulation studies (e.g., Boes et al., 2018), we believe that this scenario is highly implausible. More likely, many of the significant findings in these studies might derive from non-optimal research practices during the data analysis, collection, and reporting process, such as running multiple types of analyses, or trying different criteria for outlier exclusion and data trimming, but only reporting those cases which yield significant results (John et al., 2012; Munafò et al., 2017; Simmons et al. 2011).

In this line, we already found some signs in agreement with the latter interpretation in SS22. For instance, 42 out of the 43 included studies were not preregistered, leaving the degrees of experimenter freedom untouched, and making more difficult for other researchers to evaluate whether the authors of a paper have engaged in non-optimal practices (Hardwicke & Wagenmakers, 2023; Lakens, 2019). Relatedly, small sample sizes (as is the case) have also been identified as hotbeds for false-positive results, since the use of non-optimal practices have stronger effects when the sample is small (Simmons et al., 2011). Moreover, several studies concluded in support of their hypotheses even when the crucial statistical contrasts were not tested or came out non-significant (see SS22 for details). Yet, it is important to clarify that our point is not to accuse the researchers in this literature of carrying out non-optimal practices knowing that they are non-optimal, or even getting involved in scientific fraud. Many researchers were until relatively recently unaware that these practices were non-optimal. Indeed, many of the practices that are currently identified as non-optimal have been historically justified as part of standard research practice (Bem, 2003; John et al., 2012). For that reason, we believe that the present paper is especially relevant to raise awareness in future researchers about these issues and encourage them to embrace more robust and transparent methods.

In a similar vein, we want to point out that present results should not be taken as evidence against the embodiment of language. As discussed in SS22, we are not directly testing whether the motor system is engaged in language comprehension, but whether the statistical information reported in these neurostimulation studies corresponds with the expectations of a reliable literature or not (Simonsohn

et al., 2014). Robust and replicable evidence is a prerequisite for concluding in favor or against any scientific theory, and here we are showing that the brain stimulation evidence published to date is not reliable enough. We thereby recommend caution before drawing any conclusion regarding the role of the motor system in language comprehension from these studies.

It is also worth noting that present results not only are in line with SS22, but also align with a growing number of studies, including preregistered and well-powered replication attempts (Montero-Melis et al., 2022; Morey et al., 2022), reanalyses of previously published work (Papesh, 2015; Witt et al., 2020), as well as other meta-analytic works (Winter et al., 2022), raising concern about key findings for the embodied view of language processing. Accordingly, we believe that future studies on this topic should urgently adopt practices such as conducting well-powered studies (e.g., Lakens, 2022b), replicating previous findings (e.g., Zwaan et al., 2018), and implementing preregistrations (e.g., Hardwicke & Wagenmakers, 2023) to generate valuable and interpretable findings on the embodied nature of language (see SS22 for a more detailed discussion; see also Solana, 2023).

Two are the main limitations of the present study. The first one relates to the number of values entered in the analysis. In their simulations, Schimmack and colleagues (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020) employed k = 100 values and larger. In the present work, we used k = 38 for the main analysis and k = 40 for the robustness analysis. The second limitation is that an important percentage of the initially selected contrasts could not be finally included in the analyses since they were not reported or reported incompletely (impeding their recalculation), and the authors of the respective papers could not or did not provide them when requested. Both limitations might affect the numerical estimations reported in the present study. We cannot estimate their influence on our results, but we believe that they do not critically affect our main conclusions, as the differences between the EDRs and the ODRs are clearly evident, and they seem to be robust across the different analyses that we performed, even when we manually included non-significant results that could not be recalculated. Nonetheless, readers should keep in mind that the number of contrasts included in the present study depends on (1) the number studies available when we conducted the literature search, (2) the quality of the reporting of the statistical information in them, and (3) the availability of the data from the original authors. Future meta-analytic studies within this literature will no doubt benefit from the publication of more studies, as well as better reporting practices and the implementation of open practices such as data sharing.

In conclusion, with an unacceptably low estimated power (≈ 20%) and clear signs of a large publication bias, the present reanalysis reaffirms the conclusion of our previous paper (Solana & Santiago, 2022): motor cortex stimulation studies of embodied language comprehension do not stand on solid ground.

We recommend taking their findings with caution and implementing more robust and transparent practices in future studies.

Acknowledgments

This work was supported by a FPU predoctoral grant (ref. FPU20/01946) to PS and by the Project PY20_00689 from the Andalusian Government and FEDER to JS (PI). The present article is part of the PhD dissertation of PS at the Psychology Doctoral Program of the University of Granada under the supervision of JS.

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Author contributions

PS: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft. JS: Supervision, Writing – Review & Editing.

Data availability

Raw data, analysis scripts, high quality plots, and supplementary information of the present study are available at the following Open Science Framework (OSF) repository: https://osf.io/4f6em/

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577-660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bartoš, F., & Schimmack, U. (2020). *zcurve: An R Package for Fitting Z-curves*. R package version 2.3.0. https://CRAN.R-project.org/package=zcurve

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology, 6*, MP.2021.2720. https://doi.org/10.15626/MP.2021.2720

Bem, D. J. (2003) Writing the empirical journal article. In J. M. Darley, M. P. Zanna & H. L. Roediger (Eds.), *The compleat academic: A career guide* (2nd Edition). American Psychological Association.

Birba, A., Vitale, F., Padrón, I., Dottori, M., de Vega, M., Zimerman, M., ... & García, A. M. (2020). Electrifying discourse: Anodal tDCS of the primary motor cortex selectively reduces action appraisal in naturalistic narratives. *Cortex, 132*, 460-472. https://doi.org/10.1016/j.cortex.2020.08.005

Boes, A. D., Kelly, M. S., Trapp, N. T., Stern, A. P., Press, D. Z., & Pascual-Leone, A. (2018). Noninvasive brain stimulation: challenges and opportunities for a new clinical specialty. *The Journal of Neuropsychiatry and Clinical Neurosciences, 30*(3), 173-179. https://doi.org/10.1176/appi.neuropsych.17110262

Boulenger, V., Mechtouff, L., Thobois, S., Broussolle, E., Jeannerod, M., & Nazir, T. A. (2008). Word processing in Parkinson's disease is impaired for action verbs but not for concrete nouns. *Neuropsychologia, 46*(2), 743-756. https://doi.org/10.1016/j.neuropsychologia.2007.10.007

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology, 4*, MP.2018.874. https://doi.org/10.15626/MP.2018.874

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Cognitive Brain Research, 24*(3), 355-363. https://doi.org/10.1016/j.cogbrainres.2005.02.020

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. https://doi.org/10.1038/nrn3475

Glenberg, A. M., Sato, M., & Cattaneo, L. (2008). Use-induced motor plasticity affects the processing of abstract and concrete language. *Current Biology, 18*(7), R290-R291. https://doi.org/10.1016/j.cub.2008.02.036

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour, 7*(1), 15-26. https://doi.org/10.1038/s41562-022-01497-2

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. https://doi.org/10.1177/1740774507079441

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Lakens, D. (2019). The value of preregistration for psychological science: a conceptual analysis. *Japanese Psychological Review, 62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2022a). Bias detection. In D. Lakens (Ed.), *Improving your statistical inferences*. https://lakens.github.io/statistical_inferences/

Lakens, D. (2022b). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris, 102*(1-3), 59-70. https://doi.org/10.1016/j.jphysparis.2008.03.004

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2020). Average power: A cautionary note. *Advances in Methods and Practices in Psychological Science, 3*(2), 185-199. https://doi.org/10.1177/2515245920902370

Montero-Melis, G., Van Paridon, J., Ostarek, M., & Bylund, E. (2022). No evidence for embodiment: The motor system is not needed to keep action verbs in working memory. *Cortex, 150*, 108-125. https://doi.org/10.1016/j.cortex.2022.02.006

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., … & Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review, 29*, 613-626. https://doi.org/10.3758/s13423-021-01927-8

Papesh, M. H. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General, 144*(6), e116. https://doi.org/10.1037/xge0000125

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*(7), 576-582. https://doi.org/10.1038/nrn1706

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Repetto, C., Colombo, B., Cipresso, P., & Riva, G. (2013). The effects of rTMS over the primary motor cortex: The link between action and language. *Neuropsychologia, 51*(1), 8-13. https://doi.org/10.1016/j.neuropsychologia.2012.11.001

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift für Psychologie / Journal of Psychology, 227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Simmons, J., Nelson, L., & Simonsohn, U. (2018, January 8). *P-curve handles heterogeneity just fine*. DataColada. http://datacolada.org/67

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. https://doi.org/10.1037/a0033242

Solana, P. (2023). The embodiment and disembodiment of language. *Nature Reviews Psychology, 2*, 391. https://doi.org/10.1038/s44159-023-00203-x

Solana, P., & Santiago, J. (2022). Does the involvement of motor cortex in embodied language comprehension stand on solid ground? A *p*-curve analysis and test for excess significance of the TMS and tDCS evidence. *Neuroscience and Biobehavioral Reviews, 141*, 104834. https://doi.org/10.1016/j.neubiorev.2022.104834

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science, 11*(5), 713-729. https://doi.org/10.1177/1745691616650874

Vitale, F., Padrón, I., Avenanti, A., & de Vega, M. (2021). Enhancing motor brain activity improves memory for action language: A tDCS study. *Cerebral Cortex, 31*(3), 1569-1581. https://doi.org/10.1093/cercor/bhaa309

Vukovic, N., Feurra, M., Shpektor, A., Myachykov, A., & Shtyrov, Y. (2017). Primary motor cortex functionally contributes to language comprehension: An online rTMS study. *Neuropsychologia, 96*, 222-229. https://doi.org/10.1016/j.neuropsychologia.2017.01.025

Winter, A., Dudschig, C., Miller, J., Ulrich, R., & Kaup, B. (2022). The action-sentence compatibility effect (ACE): Meta-analysis of a benchmark finding for embodiment. *Acta Psychologica, 230*, 103712. https://doi.org/10.1016/j.actpsy.2022.103712

Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. C. (2020). Reanalysis Suggests Evidence for Motor Simulation in Naming Tools Is Limited: A Commentary on Witt, Kemmerer, Linkenauger, and Culham (2010). *Psychological Science, 31*(8), 1036-1039. https://doi.org/10.1177/0956797620940555

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. https://doi:10.1017/S0140525X17001972