This is the accepted version of

Solana, P., and Santiago, J. (in press). 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.

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

Pablo Solana[a,b,*] and Julio Santiago[a,b]

[a] Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain

[b] Department of Experimental Psychology, University of Granada, Granada, Spain

* Corresponding author at: Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Campus of Cartuja, 18011 Granada, Spain.

E-mail address: solana@ugr.es (Pablo Solana).

Abstract

According to the embodied cognition view, comprehending action-related language requires the participation of sensorimotor processes. A now sizeable literature has tested this proposal by stimulating (with TMS or tDCS) motor brain areas during the comprehension of action language. To assess the evidential value of this body of research, we exhaustively searched the literature and submitted the relevant studies ($N = 43$) to *p*-curve analysis. While most published studies concluded in support of the embodiment hypothesis, our results suggest that we cannot yet assert beyond reasonable doubt that they explore real effects. We also found that these studies are quite underpowered (estimated power < 30%), which means that a large percentage of them would not replicate if repeated identically. Additional tests for excess significance show signs of publication bias within this literature. In sum, extant brain stimulation studies testing the grounding of action language in the motor cortex do not stand on solid ground. We provide recommendations that will be important for future research on this topic.

***Keywords**:* Embodied cognition, Language comprehension, Motor system, TMS, tDCS, *P*-curve analysis, Excess significance.

1. Introduction

How meaning is represented in the mind and the brain is a topic at the heart of cognitive science. According to classic cognitive theories (e.g., Fodor, 1975), meaning is represented by means of abstract and arbitrary symbols that rely on higher-order, amodal brain regions (e.g., Mahon & Caramazza, 2008). On the contrary, embodied cognition theories propose that conceptual knowledge is grounded on the sensorimotor experience we accumulate through our interaction with real-world referents. Under this view, language understanding is mediated by detailed mental simulations produced by the (re)activation of modality-specific brain areas primarily involved in action, perception, and emotion (for reviews, see Barsalou et al., 2008; Binder & Desai, 2011; Fischer & Zwaan, 2008; Kiefer & Pulvermüller, 2012; Meteyard et al., 2012).

A prominent and fruitful line of research within the embodied semantics framework has focused on the role that the motor system plays in representing the meaning of action-related language (e.g., action verbs or manipulable object nouns). For instance, in their seminal study, Glenberg and Kaschak (2002) found that processing phrases describing movements in a specific direction (e.g., "close the drawer") facilitates motor responses in the same direction (i.e., away from your body). This effect was named the Action-Sentence Compatibility Effect (ACE) and has been reported in many other studies (e.g., de Vega et al., 2013; Glenberg et al., 2008; but see Morey et al., 2022). Another relevant source of evidence comes from neuroimaging studies, which have consistently shown that the processing of action verbs (e.g., "pick" or "kick") activates the same motor areas involved in the execution of hand and foot movements, respectively (e.g., Aziz-Zadeh et al., 2006; Hauk et al., 2004; Tettamanti et al., 2005).

Nevertheless, a pending challenge for the embodied language framework is to convincingly demonstrate that the motor system has a truly functional role in meaning

representation and, therefore, that it plays a core role in semantic processing, instead of being a mere correlate or the result of post-conceptual processes (Mahon & Caramazza, 2008; Ostarek & Bottini, 2021; Ostarek & Huettig, 2019). As a direct response to this caveat, there is a growing body of studies that use neurostimulation techniques such as Transcranial Magnetic Stimulation (TMS; Walsh & Cowey, 2000) and Transcranial Direct Current Stimulation (tDCS; Nitsche & Paulus, 2000) to shed light on this controversial issue (for a review of TMS studies of embodied language processing, see Papeo et al., 2013).

These techniques are based on the application of non-invasive stimulation over the scalp (magnetic pulses in the case of TMS and electric current in the case of tDCS) to temporally modulate the underlying neural activity and, in consequence, alter cognition, behavior, and physiological activity (for an overview, see Polanía et al., 2018). These methods have clear advantages over more classic neuroscientific strategies (e.g., lesion-based studies) when assessing functional links between brain and behavior, mainly because of the non-invasive and transitory nature of the neural modulation.

The application of TMS pulses over the motor cortex leads to fast and effector-specific modulations of muscle excitability, measured through motor evoked potentials (MEPs; Bestmann & Krakauer, 2015). Several studies have shown that processing action-related language that implies the same effector being stimulated affects the size of MEPs in early temporal windows (e.g., Buccino et al., 2005; Gianelli & Dalla Volta, 2015; Glenberg et al., 2008; Innocenti et al., 2014; Labruna et al., 2011), evidencing the participation of the motor cortical areas in accessing the meaning of language.

Brain stimulation can also be used to test the causal contribution of a brain region in a particular cognitive process (Polanía et al., 2018). In this line, the accumulated evidence demonstrates that disturbing the motor or premotor cortex activity by using repetitive TMS (rTMS; e.g., Lo Gerfo et al., 2008; Repetto et al., 2013; Vukovic et al., 2017; Willems et al.,

2011) or tDCS (e.g., Birba et al., 2020; Gijssels et al., 2018; Niccolai et al., 2017; Vitale et al., 2021) alters the comprehension of action-related language but not the processing of non-action language (e.g., abstract verbs; but see Johari et al., 2021). This pattern of findings strongly suggests a causal role for the motor system in action language understanding.

Thereby, the impression that one can obtain from this literature is that extant TMS and tDCS studies support the grounding of action language in the motor cortex (see the full set of currently published studies in Table 1: out of 43 studies, 40 support embodiment). Yet, not all the available evidence points out in this direction. For example, Papeo et al. (2009) showed that the modulation of the hand MEPs during action language processing only occurred if the TMS pulses were applied in advanced processing stages (i.e., 500 ms post-stimulus). Similarly, Tomasino et al. (2008) found that the application of TMS over the primary motor cortex affected action verb processing when participants were explicitly asked to generate a motor simulation of the action but not during purely linguistic tasks such as silent reading or frequency judgment. These results suggest that the motor recruitment observed during language processing may not be an integral part of meaning construction (Mahon & Caramazza, 2008; Papeo et al., 2015).

Furthermore, the studies that do report results congruent with the embodied view are not always consistent in their findings (Ostarek & Bottini, 2021; Ostarek & Huettig, 2019; Shebani & Pulvermüller, 2018; Togato et al., 2021). For instance, while some studies observe facilitatory interactions between motor activation and language processing (e.g., lower reaction times or higher MEPs amplitudes; e.g., Pulvermüller et al., 2005; Gianelli & Dalla Volta, 2015; Vicario & Rumiatti, 2012; Willems et al., 2011), other studies report inhibitory interactions (e.g., higher reactions times or lower MEPs amplitudes; e.g., Buccino et al., 2005; Candidi et al., 2010b; Kuipers et al., 2013; Vukovic et al., 2017).

These inconsistencies could be explained by differences between the studies, such as the stimulation protocol or the kind of stimuli employed (Boulenger et al., 2006; Innocenti et al., 2014). Moreover, it is important to keep in mind that behavioral evidence points out the importance of conceiving of motor-language interactions as dynamic and flexible processes (Boulenger et al., 2006; de Vega et al., 2013; Shebani & Pulvermüller, 2018; Togato et al., 2021). Nevertheless, some results are hard to reconcile. Probably, the best example is Gianelli and Dalla Volta's (2015) study, which is a better-powered replication of Buccino et al.'s (2005) influential study. Buccino et al. (2005) found that listening to hand and foot-related sentences decreased MEPs amplitudes for hand and foot muscles, respectively. On the contrary, Gianelli and Dalla Volta (2015) found in their replication an increase in hand MEPs amplitudes when processing hand-related phrases and no effect whatsoever for foot MEPs.

Inconsistent findings such as these ones open one possibility that has remained unexplored so far: Could these inconsistencies be caused by an accumulation of false-positive findings, as the result of low statistical power and sub-optimal research practices? And if so, how valid and reliable are in fact the results of this body of research?

It is now widely accepted that the psychological and neuroscientific literature suffers from a credibility and replicability crisis (Button et al., 2013; Open Science Collaboration, 2015; Simmons et al., 2011; although the problem is also present in many other disciplines, such as cancer biology; e.g., Errington et al., 2021). Indeed, it is estimated that the replicability rate in psychological science is less than 40% (Open Science Collaboration, 2015) and the statistical power underlying neuroscience is around 20% (Button et al., 2013). Its causes are multiple, but we can summarize them into three main ones. In the first place, there is the use of underpowered studies, mainly because of small sample sizes, that increase the probability of obtaining false negatives (Type II error) and inflated true effect sizes when a positive result is obtained (Button et al., 2013; Ioannidis, 2005). Secondly, flexibility in

data analysis and collection (e.g., in criteria for dropping outliers or running multiple types of analyses), commonly known as *p*-hacking (Simmons et al., 2011), usually renders some significant tests that favor the researchers' hypotheses, generating the illusion that there is an effect when it is really absent (i.e., false positives or Type I error). This flexibility interacts with flexibility in the establishment of hypotheses, as in some cases hypotheses are formulated after the results are known (i.e., HARKing; Kerr, 1998). Finally, there is publication bias, namely the selective publishing of significant results, while non-significant ones remain hidden (i.e., the file drawer problem; Rosenthal, 1979), giving rise to a literature that gives a misleading impression about a certain topic.

Results in line with this crisis have recently started to come out in the embodied semantics literature. For instance, a preregistered, multi-lab study (Morey et al., 2022) failed to replicate the influential Action-Sentence Compatibility Effect (ACE; Glenberg & Kaschak, 2002) across 18 labs. Similarly, in a highly-cited article, Witt et al. (2010) concluded that the motor system plays a crucial role in grounding the meaning of manipulable objects. However, a recent reanalysis using multiverse analyses and Bayesian statistics (Witt et al., 2020) showed that the results of Witt et al. (2010) did not contain enough evidence for supporting their conclusions and that the original results depended upon decisions that researchers made during data analysis (e.g., outlier exclusion or the type of statistical analysis).

Despite the inconsistent results and the shadow of the replicability crisis, there are no studies that have set out to quantitatively assess the reliability of neurostimulation (TMS and tDCS) studies of embodied language comprehension. In the present study, we aimed to evaluate the evidential value, the underlying statistical power, and the presence of publication bias in the available studies of this literature by using *p*-curve analyses (Simonsohn et al., 2014a, 2014b, 2015) complemented with excess significance tests (Ioannidis & Trikalinos, 2007; Ioannidis, 2013).

P-curve analysis is a novel meta-analytic tool that operates based on how the significant *p*-values ($p < 0.05$) of a set of studies are distributed (Simonsohn et al., 2014a). When the null hypothesis is true (i.e., there are no underlying true effects), all *p*-values are expected to be equally likely and, therefore, to be distributed uniformly over the whole range between 0 and 1. Importantly, that also applies in the range between 0 and 0.05, which is the range of published significant findings. On the contrary, if there are true effects underlying a set of studies and the studies have enough power to detect them, the distribution of significant *p*-values is expected to be right-skewed, meaning that *p*-values close to 0 (e.g., $p = 0.0001$) should be more likely than *p*-values close to the conventional significance threshold of $p = 0.05$ (e.g., $p = 0.045$). Therefore, the degree of right skewness in the distribution of *p*-values between 0 and 0.05 gives us an indication of whether there are true effects underlying the set of studies introduced in the analysis and the power of those studies to detect such effects. If there are true underlying effects, the right-skewness of the distribution should increase the greater the statistical power of the set of studies (Simonsohn et al., 2014a, 2014b). On the contrary, finding a left-skewed distribution, where *p*-values close to 0.05 are more likely than those close to 0, would be an indication of the researchers' attempts for obtaining significant results (below the established $p = 0.05$) when the null hypothesis is true; or, in other words, the existence of severe *p*-hacking in the literature (for examples of published studies applying *p*-curve analyses to other controversial topics such as power posing or ego depletion, see Simmons & Simonsohn, 2017; Vadillo et al., 2016a).

We decided to use *p*-curve analysis because it allows us to deal with some of the limitations of other meta-analytic methods (e.g., the "trim and fill" method; Duval & Tweedie, 2000). First of all, it is important to point out that our target literature includes several sources of heterogeneity, like the stimulation method (e.g., TMS vs. tDCS), the nature of the dependent variables (e.g., behavioral vs. physiological) or the kind of linguistic stimuli

(e.g., action verbs vs. manipulable object nouns). Contrary to other meta-analytic methods, the outcomes provided by *p*-curve analyses rest upon statistical principles (i.e., how *p*-values distribute) that are independent of the degree of homogeneity of the set of studies (Simonsohn et al., 2014a; see also Simmons et al., 2018), making *p*-curve analysis an ideal tool for analyzing heterogeneous bodies of research. Furthermore, because *p*-curve analysis only includes significant values, its results are protected against publication biases (i.e., the non-publication of negative or null findings; Simonsohn et al., 2014a).

Nonetheless, *p*-curve analysis does not directly test for publication bias (although observing a clear left-skewed distribution suggests evidence for severe *p*-hacking; Simonsohn et al., 2014a). For this reason, to evaluate the presence of biases in this literature, we also run tests for excess significance (TES; Ioannidis & Trikalinos, 2007). This method allows comparing the observed proportion of statistically significant results in a body of findings with its expected proportion if there is a real underlying effect (i.e., its expected statistical power). In case the observed proportion is significantly greater than the expected one, then the findings derived from a set of studies are considered "too good to be true" (Francis, 2012), thus suggesting the presence of publication bias.

2. Method

2.1. Transparency and openness

For the sake of the transparency and reproducibility of the present work, detailed tables of the study selection process, a disclosure table of the source of the *p*-values included in the analysis, raw data, analysis scripts, and complementary analyses can be found in the Supplementary Material section and on the Open Science Framework (OSF) webpage: https://osf.io/ehcga/?view_only=8e7765ddc74f4893abbd97cf6f1e4bb8

2.2. Eligibility criteria

We limited the present review to studies that met the following criteria: (1) being a peer-reviewed research article published in an international journal in English; (2) employing either TMS or tDCS for modulating the activity of the brain motor areas; and (3) studying the grounding of action-related language in the motor system.

As we are interested in meaning construction during the processing of action-related concepts, not only "classic" language comprehension studies were eligible (e.g., lexical decision studies as Willems et al., 2011), but also memory (e.g., Vitale et al., 2021) or learning (e.g., Liuzzi et al., 2010) studies which experimental tasks required the semantic processing of action-related language. Because the analysis unit is the experiment, studies containing multiple experiments were eligible as long as they included at least one experiment that met our inclusion criteria. Only suitable experiments were taken into account for the analyses.

2.3. Literature search and study selection

In order to localize all the neurostimulation studies within the field of embodied language comprehension, we structured our literature search in three different stages based on the PRISMA guidelines for conducting systematic reviews and meta-analyses (Moher et al., 2009; Page et al., 2021).

In the first stage, we searched the Web of Science and Scopus databases by means of the following query: ("motor system" OR "motor cortex" OR "premotor cortex") AND ("language comprehension" OR "language understanding" OR "language processing" OR "semantic processing") AND ("TMS" OR "rTMS" OR "tDCS"). The latest search was made in May 2021. This procedure allowed us to identify 165 potential studies (76 from Web of Science and 89 from Scopus), that turned into 133 records after removing 32 duplicates.

Titles and abstracts of these 133 studies were screened, and those that did not fulfill our inclusion criteria were removed as follows.

First, we eliminated 86 articles that had nothing to do with the topic of embodied semantics. This includes studies of motor cortex neurostimulation during language processing with goals other than evaluating the meaning construction of action-related concepts (e.g., studies derived from the motor theory of speech perception aimed at testing whether the motor cortex is involved in speech perception; e.g., Schomers et al., 2015). Second, we discarded six papers related to the topic but which were not peer-reviewed empirical studies (i.e., reviews, meta-analyses, book chapters, or conference papers). Finally, we discarded four empirical studies of embodied language comprehension that were not TMS or tDCS studies (e.g., fMRI or EEG studies).

The remaining 37 neurostimulation studies were assessed for eligibility by both authors of the present work. Disagreements were resolved by consensus. Given that we are interested in embodiment effects during language understanding, we decided to discard three studies that focused on gesture processing more than language comprehension (De Marco et al., 2018; Hayek et al., 2018; Murteira et al., 2018). For a similar reason, we discarded Meister et al.'s (2012) study considering that its experimental task involved processing only non-linguistic stimuli (action-related images and videos). Lastly, we also discarded Papeo et al.'s (2015) study because it is mainly centered on the interaction between the motor cortex and the left posterior middle temporal gyrus (lpMTG) and we were only interested in motor areas. Therefore, a total of 32 studies were selected at this first stage.

In the second stage of our literature search, we consulted the reference lists of those 32 studies. This procedure allowed us to identify five extra studies that met our inclusion criteria. The references of these new five papers were also checked, but no more suitable studies were located. Consequently, the article sample added up to 37 studies.

Finally, the third stage consisted in finding papers that cited any of these 37 articles. We manually entered each one of them in Google Scholar and then carefully checked the studies listed in the "Cited by" section. The latest search was carried out in October 2021. Seven extra studies were located, but one of them (Dupont et al., 2020) was discarded because it was still in a preprint version. This increased our article sample to 43 studies.

In sum, a total of 43 studies containing 47 suitable experiments were selected for the *p*-curve analysis. All articles selected are marked with an asterisk (*) in the reference list. A flowchart that synthetizes the literature search process is depicted in Figure 1. The main characteristics of the selected studies are compiled in Table 1. The complete list of studies and experiments located through the search alongside their exclusion criteria can be found in Supplementary Material, Appendix 1.

[PLEASE, PLACE FIGURE 1 NEAR HERE]

[PLEASE, PLACE TABLE 1 NEAR HERE]

2.4. Contrast selection

In order to assure the transparency and reproducibility of the present study, following the guidelines of Simonsohn et al. (2014a), we created a disclosure table that contains detailed information on each contrast we selected for the *p*-curve analysis (see Supplementary Material, Appendix 2).

We started by identifying the researchers' hypothesis of interest for each experiment. Next, we established the study design that allowed the researchers to test that hypothesis (e.g., 2 x 2 or 3 x 2). For this step, we did not consider variables that were included in the analyses of the studies but did not clearly derive from the researchers' predictions (i.e.,

variables used for exploratory purposes). After that, we identified the key statistical result that tested the stated hypothesis in the design. For simpler and more usual designs (e.g., 2 x 2 or 3 x 2), we followed the guidelines offered by Simonsohn et al. (2014a). Broadly, for attenuated interactions (i.e., when the presence of a certain variable reduces or eliminates an effect), we selected the interaction test, while for reversing interactions (i.e., when the presence of a certain variable reverses an effect), we selected the simple effects that contribute to the interaction. For example, Willems et al. (2011) expected that theta-burst TMS over the left premotor cortex (vs. the right premotor cortex) will impair language comprehension for action-related verbs but not for abstract verbs (i.e., attenuated interaction). Thus, we selected the value of the two-way interaction between Stimulation site and Type of verb. In another study, Pulvermüller et al. (2005) predicted that applying TMS pulses over the hand primary motor cortex will facilitate the processing of hand-related verbs (compared to foot-related verbs), whereas applying the pulses over the foot primary motor cortex should originate the opposite pattern (i.e., reversing interaction). Hence, we selected the two main effects that form the expected interaction between Stimulation site and Stimulus type. For studies with more complex designs (e.g., 4 x 2 or 3 x 2 x 2), for which Simonsohn et al. (2014a) do not offer clear enough guidelines, we tried to apply the same logic, always justifying our decision in the disclosure table (as it has been done in previous *p*-curve studies, e.g., Navas et al., 2021).

Finally, for each study, we extracted the value of the selected statistical test. Because in some cases statistics are reported inexactly (i.e., authors only report that the *p*-value is smaller than a benchmark; e.g., $p < 0.05$), or they are misreported (Nuijten et al., 2016), Simonsohn et al. (2014a) recommend recomputing the exact *p*-value associated with each of the selected tests. We carried out these recalculations by means of the *P*-curve App 4.06 (http://www.p-curve.com/app4/). If the test was not significant, we report it as "not

significant" in our disclosure table. When the selected contrast was not reported or it was reported in a way that impeded recalculating its respective *p*-value (e.g., authors do not report the value of the statistical test that provided the *p*-value), we emailed the corresponding author of the paper to request the missing values. We received none of the requested statistics (belonging to nine studies): in two cases, the author mentioned not having access to those data anymore and, in the other seven cases, we got no response. We report these cases (i.e., when the crucial test was not reported or it was reported incompletely) in our disclosure table as "not reported" and "missing information", respectively. Crucially, if the key test was not significant, was not reported, or was incompletely reported, under no circumstances did we include a contrast other than the selected one, since doing so increases the probability of finding evidential value in a set of studies, even though they actually lack it (Simonsohn et al., 2015).

The recalculated *p*-values from four studies (Johari et al., 2021; Labruna et al., 2011; Pulvermüller et al., 2005; Vukovic & Shtyrov, 2019) differed from the original values reported in the papers. In the cases of Pulvermüller et al.'s (2005) and Johari et al.'s (2021) studies, discrepancies arose from the internal functioning of the app: while the original authors applied one-sided tests, the app always recomputes all tests as two-sided. Regarding Vukovic and Shtyrov's (2019) study, after contacting the authors of the paper, the discrepancy was tracked down to errors in the reporting of the degrees of freedom of their statistics. Corrected results were included in the disclosure table and were thus used in the analysis. Nevertheless, because they applied a false discovery rate correction (FDR) that cannot be implemented in the app, the recalculated *p*-values remain still slightly different from the ones reported in the paper. Finally, we also found that the recalculated *p*-values from Labruna et al.'s (2011) study did not match the ones reported in the paper: they were reported as significant, while the recalculated *p*-values did not reach significance. These

inconsistencies could not be accounted for by any of the usual causes (e.g., one-sided tests or multiple comparison corrections). We contacted the authors, but they did not provide any explanation about this issue. As the misreporting affected the crucial tests of this study and only significant *p*-values are included in *p*-curve analyses, this effectively meant that no contrast from Labruna et al.'s (2011) study could be included in the analyses.

On many occasions, more than one contrast can be eligible for each study. For example, there are studies with more than one hypothesis, and also studies containing hypotheses that can be tested in multiple ways (e.g., a hypothesis formulated in "performance" terms can be tested both by using reaction times or accuracy measures). In these cases, in order to reduce the presence of any selection bias, Simonsohn et al. (2014a) recommend creating a rule for selecting two alternative but equally valid contrasts and using them for a "main analysis" and a "robustness analysis", respectively. (Note that despite the use of the terms "main" and "robustness", they do not indicate the preponderance of one analysis over the other. Both analyses are complementary and equally important.) As in Vadillo et al. (2016a) and Navas et al. (2021), we selected the first contrast reported in the paper for the main analysis and the second one for the robustness analysis. The same was applied for studies with multiple hypotheses: the first hypothesis formulated in the paper was used for selecting the contrast for the main analysis and the second hypothesis was used for the robustness analysis. Any deviation from these rules was justified in the disclosure table. In cases where only one possible contrast was eligible, we used the same value for both the main and the robustness analysis.

In short, for the main analysis, a total of 54 possible values were identified. However, 14 of them were not significant, three were not reported, and five were incompletely reported, impeding to recalculate their respective *p*-values. Regarding the robustness analysis, 55 potential contrasts were identified, but 14 of them were not significant, three were not

reported, and six were incompletely reported. Therefore, a total of 32 values were included in the main analysis, and 32 values were used for the robustness analysis.

2.5. Data analysis and statistical inference

P-curve analyses were conducted using the *P*-curve App 4.06 (http://www.p-curve.com/app4/). We carried out two complementary analyses: a main analysis ($k = 32$) and a robustness analysis ($k = 32$; see the disclosure table in Supplementary Material, Appendix 2). Moreover, to make sure that the results are not biased by the inclusion of those *p*-values that differed from the ones reported in the corresponding papers (see section 2.4 – Contrast selection), we carried out the analyses with and without those values (the latter are reported as additional analyses in the Supplementary Material, Appendix 4).

The first test conducted by the app is a test for right-skewness (Simonsohn et al., 2014a). This test assesses whether the distribution of the selected *p*-values is significantly more right-skewed than a flat distribution (as expected if the set of studies explores true effects versus null effects). Following Simonsohn et al. (2015), if the test for the "half *p*-curve" (i.e., values ranging from 0 to 0.025) is significant with an *alpha* of 0.05, or if the tests for both the full (i.e., values ranging from 0 to 0.05) and the half *p*-curve are significant with an *alpha* of 0.1, then the test for right-skewness suggests that the analyzed set of studies contains evidential value (i.e., the majority of them explore true effects). These criteria were created to counter "ambitious *p*-hacking" cases (e.g., trying to obtain a $p = 0.03$ instead of a $p = 0.045$) that rise the probability of declaring that a set of studies contains real effects even when they are not actually present (Ulrich & Miller, 2015). Focusing on the half *p*-curve is supposed to exclude a large percentage of these cases, thus providing a less biased evaluation of the evidential value (Simonsohn et al., 2015).

When the *p*-curve is not significantly right-skewed, the next step is to carry out a test for flatness (Simonsohn et al., 2014a). In this case, the app assesses whether the entered distribution of *p*-values is significantly flatter (as expected if the set of studies explores null effects) than the one expected if a set of studies explores true effects but with an underlying statistical power of just 33%. If the test for flatness is statistically significant with an *alpha* of 0.05, then it suggests that the set of studies does not contain evidential value (i.e., the majority of them explore null effects). However, if both the right-skewness and the flatness tests are not significant, then the *p*-curve analysis is declared inconclusive regarding its evidential value (Simonsohn et al., 2014a).

To assess whether the results of the above-mentioned tests are reliable, we also computed a cumulative meta-analysis. It consists in plotting how the significance level of both the right-skewness test and the flatness test changes if we progressively exclude the most extreme *p*-values included in the analysis (i.e., those closer to either 0 or 0.05) until reaching half of the entered values. If the results of the tests hinge on a few extreme values, then we should not place too much confidence in them (Simmons & Simonsohn, 2017).

In addition, *p*-curve analysis also provides an estimation of the underlying statistical power of the set of studies (Simonsohn et al., 2014b; see also Simmons et al., 2018). This is computed by comparing the degree of fit between the distribution of the entered *p*-values and the expected *p*-curves for each value between 5% and 99% of power. The resultant power estimate is the value with the best fit.

Finally, to evaluate the presence of publication bias in this literature, we compared the observed proportion of statistically significant results in this body of findings with its expected underlying power by means of tests for excess significance (Ioannidis & Trikalinos, 2007). If the observed proportion is significantly greater than the expected one, then the test suggests the presence of publication bias. We used the number of significant *p*-values

included in the *p*-curve analysis vs. the number of non-significant *p*-values to compute the observed proportion of significant findings in the literature. The estimation of power provided by the *p*-curve analysis was used as the proportion of significant results that we should expect. Those values that were not reported in their respective papers were excluded from the analysis, given that we cannot be certain about whether they are significant or not. These calculations were carried out by means of one-tailed binomial tests (as in Vadillo et al., 2016b), using the *jmv* package in R (R Core Team, 2021).

All the present *p*-curve analyses can be replicated by copy-pasting the *p*-values included in the disclosure table (Supplementary Material, Appendix 2) into the *P*-curve App (http://www.p-curve.com/app4/). Tests for excess significance can be replicated by running the R script provided in the Supplementary Material, Appendix 3B with the data in Appendix 3A.

3. Results

The distribution of *p*-values for the main analysis is shown in Figure 2A. The right-skewness test was significant for both the full curve ($Z = -2.72$, $p = 0.003$) and the half curve ($Z = -1.67$, $p = 0.046$). The flatness test was not significant neither for the full curve ($Z = -0.31$, $p = 0.38$) nor for the half curve ($Z = 5.07$, $p > 0.9$). The cumulative meta-analysis for the main analysis is depicted in Figure 3A. As it shows, the significance of the right-skewness test for the full curve vanishes if only two extreme *p*-values are removed. More crucially, the same happens for the half curve if we just remove the most extreme *p*-value ($Z = -0.92$, $p = 0.18$). On the contrary, the flatness test remains non-significant independently of extreme values.

The estimated underlying statistical power for this set of values is 29% (*90% CI*: [10%, 53%]; Figure 4A). The proportion of significant results introduced in our analysis was

69.5%. The test for excess significance indicates that this proportion is significantly greater than the level of power estimated from the *p*-curve (29%; $p < 0.001$), even when considering the upper limit of its 95% confidence interval (53%; $p = 0.017$).

Regarding the robustness analysis, its *p*-curve is shown in Figure 2B. Again, the right-skewness test was significant for both the full curve ($Z = -2.47$, $p = 0.007$) and the half curve ($Z = -1.51$, $p = 0.066$), considering a significance threshold of $p < 0.1$ (see section 2.5 - Data analysis and statistical inference). Conversely, the flatness test was not significant neither for the full curve ($Z = -0.6$, $p = 0.27$) nor for the half curve ($Z = 4.87$, $p > 0.9$). The cumulative meta-analysis for the robustness analysis is shown in Figure 3B. Similar to the main analysis, the significance of the right-skewness test for the full curve only hinges on three extreme *p*-values. More importantly, the right-skewness test for the half curve relies on the inclusion of just one extreme *p*-value ($Z = -0.94$, $p = 0.17$). The flatness test remains non-significant independently of extreme values.

The estimated underlying statistical power for this set of values is 25% (*90% CI*: [9%, 49%]; Figure 4B). Again, the test for excess significance indicates that the observed proportion of significant *p*-values (69.5%) is significantly greater than this expected level of power (25%; $p < 0.001$), even if we consider the upper limit of its 95% confidence interval (49%; $p = 0.004$).

[PLEASE, PLACE FIGURE 2 NEAR HERE]

[PLEASE, PLACE FIGURE 3 NEAR HERE]

[PLEASE, PLACE FIGURE 4 NEAR HERE]

Running the analyses after excluding the *p*-values from those studies in which the original and recalculated *p*-values did not match does not significantly change the results (see Supplementary Material, Appendix 4).

4. Discussion

The present work aimed to evaluate the evidential value of the results derived from brain stimulation studies that test the predictions of the embodied semantics framework regarding the involvement of motor cortex in language comprehension (e.g., Buccino et al., 2005; Pulvermüller et al., 2005; Willems et al., 2011). To do so, we first identified all the relevant studies published up to now in international journals (N = 43). Then, we quantitatively assessed the soundness of their results by means of *p*-curve analyses (Simonsohn et al., 2014a) complemented with tests of excess of significance (Ioannidis & Trikalinos, 2007).

4.1. Do neurostimulation studies of embodied semantics stand on solid ground?

As Table 1 shows, out of 43 studies included in the analyses, only three conclude against the embodiment thesis, while the remaining 40 concluded in support. Hence, the impression that researchers can obtain from an overview of this literature is that current brain stimulation studies support, to a large extent, the grounding of meaning in the motor cortex. Akin to this impression, in a first approach, present results suggested that this set of studies contains evidential value (i.e., the majority of these studies explore true effects), as indexed by significantly right-skewed curves for both the main and the robustness analysis (Figure 2). Nonetheless, and crucially, this right skewness vanished after removing just one of the 32 values entered in the analyses (Figure 3). Following Simmons and Simonsohn (2017), if the results hinge on very few values, we should not place too much confidence in them. In the

present case, it seems obvious that we cannot thus consider that the present *p*-curves show evidential value. However, we did not find that those curves were significantly flat either (Figure 2), as would be expected if the set of studies lacks evidential value (i.e., the majority of the studies do not explore true effects). Hence, the resultant *p*-curves are not as right-skewed as expected if they have evidential value, nor flat enough to conclude that they lack it. Therefore, accordingly to the guidelines of Simonsohn et al. (2014a), concerning its evidential value, we declare our analysis as inconclusive. However, this does not prevent us from extracting some important conclusions and implications for the embodied language literature.

First, an inconclusive *p*-curve analysis means that the analyzed set of values is too noisy for allowing any clear interpretation of its evidential value (Simonsohn et al., 2014a), so we cannot confirm whether these studies explore real effects or not. Second, following Simonsohn et al. (2014a), when a *p*-curve is inconclusive, more *p*-values are necessary to establish its evidential value. It is important to emphasize that we performed an exhaustive literature search that provided us with (presumably) all the studies on the topic published in international journals, and our *p*-curves include more values than other previously published *p*-curve analyses (e.g., Burns et al., 2019; Simmons & Simonsohn, 2017). Thus, a noteworthy conclusion we can draw from the analysis of this body of research is that, after almost two decades and more than 40 studies, the work published to date does not yet allow us to establish that there are real underlying effects and thus, that the hypothesis that the motor system is functionally implicated in language understanding is supported by the evidence. This is a shocking result that contrasts with the impression that readers could obtain regarding the state of the art on this topic. Therefore, our results suggest that more quality research is needed to reach clear conclusions about embodiment effects in neurostimulation studies.

Another striking result of the present study concerns the estimated underlying power of the analyzed set of studies. Specifically, our *p*-curve analysis estimates power at just 29% (*95% CI:* [10%, 53%]) for the main analysis and 25% (*95% CI*: [9%, 49%]) for the robustness analysis. These values are far from the recommended 80% power (Button et al., 2013; Cohen, 1988), suggesting that these studies are underpowered. This level of power means that less than 30% of them would be significant if they were repeated (up to around 50% in the best scenario; see Simmons & Simonsohn, 2017), which poses important replicability issues for this set of findings. However, readers should keep in mind that this conclusion only applies if the studies were repeated exactly as the original (i.e., same sample, stimuli, procedure, and analysis).

The low replication rate predicted by the analysis may also be taken to suggest that the majority (around 70%) of the findings derived from these neurostimulation studies of embodied semantics are false positives. Nevertheless, this affirmation should be treated cautiously since the relation between power and false positives is complex and not always intuitive. Statistical power is the probability of finding a significant effect when, in fact, there is an underlying effect. The false-positive rate is the probability of obtaining a significant result when there is no underlying effect (Type I error). These two probabilities are, in principle, independent: a low power increases the false-negative rate (Type II error) but does not affect the false-positive rate (Button et al., 2013; Ioannidis, 2005). For this reason, the analyzed studies might all be reporting true findings with a low power (e.g., 30%). In this case, conducting these experiments again using a highly-powered design could lead to a replication rate of 100%. However, another plausible scenario is that this body of research contains both studies exploring true effects and studies reporting false positives. To illustrate, imagine that half of the studies report true effects with a power of 55%, while the other half report false-positive results. Averaging 55% and 5% - which is the proportion of significant

results expected by chance (i.e., false positives) with an *alpha* of 0.05 - will result in a computed power of 30%. *P*-curve analysis does not allow us to discern between these possibilities, but what is clear is that low power reduces research outcomes' value, making science uninterpretable, non-replicable, and inefficient (Button et al., 2013).

Does this body of research contain traces of publication bias? The answer seems to be yes. Tests for excess significance show that the observed proportion of significant *p*-values among these studies (around 70%) is way higher than the expected proportion according to our power estimation (around 30%, up to around 50% if we take the upper limit of its confidence interval). This result suggests that the findings in this literature are "too good to be true" (Francis, 2012). In other words, this literature contains signs of publication bias (Ioannidis & Trikalinos, 2007). However, as discussed by Ioannidis (2013), tests for excess significance do not inform us about what specific practices lie beneath the bias (e.g., selective publication of significant findings, *p*-hacking, or even scientific fraud). In the present case, considering the low replication rate predicted by our results, it is plausible that some teams had tried and failed to conceptually replicate the findings of this body of research and that, given the reticence of both researchers and scientific journals to publish null results (Franco et al., 2014), many of these studies have had trouble coming to light (i.e., a file drawer problem; Rosenthal, 1979). Whatever the case, the presence of publication bias clearly undermines the confidence we can place in this set of studies.

In a nutshell, these results suggest that this literature is not as reliable as it could seem, calling for the urgent implementation of several methodological changes.

4.2. Recommendations for future studies

First of all, we believe that the shortage of power in these studies can be traced mainly to their sample sizes. With few exceptions (e.g., Birba et al., 2020; Gijssels et al., 2018;

Liuzzi et al., 2010; Vukovic & Shtyrov, 2019), the majority of these investigations use quite small sample sizes (M_N = 22.7, SD_N = 16.4; see Figure 5 and Table 1), which is associated with low statistical power and, in consequence, with the propensity to find false negatives and overestimate true effect sizes (Button et al., 2013; Ioannidis, 2005). Moreover, small samples are more likely to be affected by the researcher degrees of freedom (e.g., trying different criteria for outlier exclusion, data trimming, and so on), which increases the probability of obtaining significant results by chance (Simmons et al., 2011). Consider the study by Gianelli and Dalla Volta (2015). In their attempt to replicate Buccino et al.'s (2005) study, they run a power analysis that indicated that, to find an effect of that size with adequate power, they would need a sample size approximately 2.5 times larger than the one employed in the original study (N = 8). Even using a more adequate sample size (N = 21), they failed to replicate Buccino et al.'s (2005) pattern of results. Consequently, one of the main conclusions that follows from the present study is that future investigations on this topic should establish sample sizes that secure enough power (i.e., using a priori power analyses; for details, see Lakens, 2022).

[PLEASE, PLACE FIGURE 5 NEAR HERE]

We would also like to point out that, except for Gianelli and Dalla Volta (2015), none of the analyzed studies has been (at least, explicitly) preregistered. Preregistration does not automatically make an investigation better, but it is a practice that prevents *p*-hacking and HARKing (Simmons et al., 2021) and also allows other researchers to evaluate science in a more transparent way (Lakens, 2019). Therefore, another recommendation we make for future work in this field is to preregister hypotheses, sample size, analysis plan, and any other experimenter degrees of freedom before carrying out data collection (for details, see

Simmons et al., 2021), which will control for indiscriminate flexibility and facilitate a less-biased evaluation of outcomes (Lakens, 2019; Simmons et al., 2021). In addition, researchers should also consider publishing their investigations as registered reports: a novel publishing model in which the article is accepted before data collection and analysis, provided that it fulfills the required quality standards (for details, see Chambers & Tzavella, 2022). This facilitates the dissemination of negative and null results, thus preventing the file drawer problem and other kinds of publication bias.

Closely related to the prior points, another key recommendation for future work is to carry out well-powered direct replications of previous findings (Nosek et al., 2022; Zwaan et al., 2018). To date, the only published direct replication of a neurostimulation study of embodied language comprehension is Gianelli and Dalla Volta's (2015) study. Considering the low replication rate predicted by the present *p*-curve analysis, a preregistered, multi-lab, neurostimulation study for testing the functional implication of the motor system in action-language understanding would be an extremely valuable piece of information (for a similar proposal, see Ostarek & Huettig, 2019).

Another important point to discuss is misreporting. When we recalculated the exact *p*-values from the key contrasts that we had selected for our analysis, we found mismatches in four studies (see section 2.4 - Contrast Selection). In two cases (Johari et al., 2021; Pulvermüller et al., 2005), those mismatches could be traced down to the use of one-sided tests, that cannot be run by the *P*-curve App. In another case (Vukovic & Shtyrov, 2019), the mismatches were due to the combination of human error (when reporting degrees of freedom) and the use of statistical corrections that cannot be implemented in our analyses, but they were without consequences for the interpretation of findings (*p*-values were correctly reported in the paper). In the last of them (Labruna et al., 2011), the misreporting (either accidental or intentional) could not be traced down to any of the usual causes and,

importantly, it affected the main conclusions of the paper, as the recalculated *p*-values did not support the embodiment thesis. The point we want to make here is that many (if not all) of these discrepancies could have been easily detected at the review stage through automated detection procedures (such as StatCheck; Nuijten et al., 2016; this tool can be accessed online at http://statcheck.io/). We contend that all scientific journals should implement these procedures as part of their routine quality checks.

A related topic is the non-reporting and the misuse of important statistical information. When we selected the key statistical contrast of each study following the guidelines of Simonsohn et al. (2014a), we noticed that a large percentage of them could not be included in the analyses because they were not significant, they were incompletely reported, or they were not reported at all (see section 2.4 - Contrast selection). Indeed, from the 43 selected studies (containing 47 experiments), we selected 54 contrasts for the main analysis and 55 contrasts for the robustness analysis, but, in both cases, only 32 of them could be included in the analyses (59.3% for the main analysis and 59.2% for the robustness analysis). For example, Tremblay et al. (2012) hypothesized that applying rTMS over the premotor cortex (vs. sham rTMS) should prevent semantic priming for action and manipulable object phrases but not for non-manipulable object and orofacial phrases. Although one-tailed *t*-tests confirmed this pattern of results, the crucial two-way interaction between Stimulation type and Phrase type was not significant. In another study, Cattaneo et al. (2010) expected that TMS over the left ventral premotor cortex (compared to TMS over the left dorsal premotor cortex and No TMS) would impair semantic processing for tool nouns but not for animal nouns. This effect was expected to arise only in congruent trials (vs. incongruent trials). Again, pairwise comparisons supported the predictions of the researchers, but neither the key three-way interaction between Stimulation type, Noun type, and Trial type nor the two-way interaction between Stimulation type and Noun type were reported in the

paper. This kind of result is interpreted in favor of the researchers' hypotheses and thus, as support for the embodied view. Nonetheless, according to some authors (e.g., Gelman & Stern, 2006; Nieuwenhuis et al., 2011), this way of analyzing interactions is incorrect and poses an important problem for the statistical validity of psychological and neuroscientific research. The presence of these practices in the brain stimulation studies of embodied language comprehension can be taken by itself as proof that their conclusions do not stand on solid ground. Thereby, practices like these should be avoided by authors and discouraged by scientific journals. As an additional note, reporting *p*-values without the value of their respective statistical tests and associated degrees of freedom, as it occurs in some other cases (e.g., Candidi et al., 2010b; Gough et al., 2013), is a practice that stands in the way of future meta-analytic work and should be avoided too.

Continuing with the statistical recommendations, we also noticed that the vast majority of the reviewed studies analyzed their data by means of analyses of variance (ANOVA). In psycholinguistic and neurolinguistic research, participants are commonly presented with lists of linguistic stimuli and the researchers intend to generalize their results both to the participants and the items population. Hence, both participants and items are random factors, since their levels are drawn by random sampling from a population. The widespread ANOVA does not allow the incorporation of more than one random factor. For this reason, data are normally averaged over the other random factor before entering it into the analysis. The main problem with this is that, by averaging over the other random factor, the analysis fails to consider part of the error variability. This inflates the probability of rejecting the null hypothesis when there is no effect (i.e., Type I error), thus leading to false-positive findings (Judd et al., 2012). One of the proposed solutions is the use of linear mixed-effects models (LMMs; for an overview, see Baayen et al., 2008). Unlike ANOVAs, the LMM approach allows taking into account the variability in the data explained by several

random factors, such as the participants and the items, thus providing more accurate and generalizable results. Future neurostimulation studies of embodied language comprehension should adopt linear mixed-effects models as a default practice in their analysis routine, which will increase the credibility of the outcomes (indeed, this practice has already been implemented in some of the extant studies in this literature; Gijssels et al., 2018; Johari et al., 2021; Niccolai et al., 2017; Monaco et al., 2021).

Another suggestive theme is how the results of the reviewed studies relate to the embodied language claims. As an example, because meaning is expected to be grounded on bodily experience, the embodiment view predicts effector-specific simulations of language content (e.g., the verb "pick" should recruit the hand motor cortex, while "kick" should recruit foot-related regions; Pulvermüller, 2005). Some brain stimulation studies have considered this aspect in their design (e.g., Buccino et al., 2005; Kuipers et al., 2013; Pulvermüller et al., 2005; Tremblay et al., 2012), but others have neglected it (e.g., Gijssels et al., 2018; Labruna et al., 2011; Oliveri et al., 2004; Willems et al., 2011). For example, Vukovic et al. (20117) assessed the effect of applying rTMS over the hand motor cortex when processing hand action verbs. As comparison condition, they used non-action, abstract verbs. Since they did not include action verbs related to other body effectors (e.g., feet or mouth), we cannot be sure that any observed implication of the motor cortex in representing language meaning is truly effector-specific. Overlooking this design aspect is problematic because it renders conclusions that can also be interpreted in agreement with theoretical accounts other than embodiment (see Chatterjee, 2010; Mahon & Caramazza, 2008). Consider the study by Onmyoji et al. (2015). They observed that reading hand movement phrases recruited the hand motor cortex, as indexed by the activation of hand muscles, measured through MEPs. In principle, this result replicates previous embodiment findings (e.g., Glenberg et al., 2008; Scorolli et al., 2012). However, sentences related to leg

movements, and even those unrelated to movement, generated the same effect on hand motor activation. And crucially, reading aloud caused more amplitude in MEPs than silent reading. These results suggest that motor recruitment during action language comprehension might be provoked by speech-related motor movements rather than by effector-specific, embodied simulations. Researchers planning to conduct future neuromodulation studies in this field should thus design their experiment in a way that allows them to confirm the embodied cognition hypotheses and, at the same time, discard alternative explanations. This not only applies to control stimuli but also to other design aspects such as the control stimulation condition or the timing of stimulation.

Finally, recent neurostimulation studies suggest functional and bidirectional links between motor and association areas outside the motor system underlying the grounding of action concepts. For instance, Papeo et al. (2015) showed that applying rTMS to the left posterior middle temporal gyrus (lpMTG) eliminated the increase in MEPs amplitude evoked by the processing of action verbs found in previous investigations (e.g., Innocenti et al., 2014; Oliveri et al., 2004). In another study, Vukovic and colleagues (2021) found that the acquisition of action-related vocabulary generates microstructural changes in prefrontal, parietal, and temporal regions and that these plasticity effects can be modulated by the application of theta-burst TMS over the motor cortex. These results are in line with current proposals arguing that mental operations rely on complex neural networks, in contrast to the idea that a certain region is selectively engaged in a particular cognitive process (for an overview, see Pessoa, 2014). Future brain stimulation studies of embodied language comprehension should also target these higher-order association regions, alongside motor system areas, in order to reach a deeper understanding of the complex neurocognitive basis of action language semantics.

4.3. Limitations

One first potential caveat that can be raised against the present study has to do with how studies and contrasts were selected for inclusion in the *p*-curve analysis. Indeed, for studies with complex designs, we needed to follow our own rules, given that Simonsohn et al. (2014a) do not offer specific guidelines for them. However, to guarantee the transparency and replicability of our results, detailed information about the implemented selection processes and all relevant decisions is provided in the Supplementary Material (see Appendixes 1 and 2), so other researchers can assess their suitability and also reanalyze our data by implementing the changes they consider appropriate.

Second, like any statistical method, *p*-curve analysis must be interpreted carefully and always considering its limitations (e.g., see Brunner & Schimmack, 2020; McShane et al., 2020). Here, we complemented our *p*-curve analysis with several additional analyses including robustness tests, cumulative meta-analyses, and tests for excess significance. Nonetheless, we encourage future researchers to complement the present findings by means of other meta-analytic methods such as "trim and fill" meta-analyses (Duval & Tweedie, 2000) or *z*-curves (Brunner & Schimmack, 2020).

4.4. Conclusions

The present study suggests that the extant brain stimulation studies that assess the grounding of action-related language in the motor system do not stand on solid ground.

First, the evidential value of these studies is unclear, so we cannot assert that they examine true effects. Second, their estimated underlying power is low (less than 30%), making the majority of them (more than 50%) not replicable in the future if identical repetitions were carried out. It is possible that some of them explore true effects (with low power), while others may be reporting false-positive findings. Third, the observed proportion

of significant results is clearly greater than what can be expected given the overall power. The main causes of this situation are, probably, small sample sizes and publication bias. In addition, we also identified other issues that contribute to the problem and leave room for improvement, such as the absence of preregistrations, cases of non-reporting and misreporting of important statistical information, and failure to adopt analyses that allow taking into account more than one random factor.

This conclusion is congruent with recently published studies that have also laid bare the fragility of the results derived from other lines of embodied semantics research (e.g., Montero-Melis et al., 2022; Morey et al., 2022; Papesh, 2015; Saccone et al., 2021; Witt et al., 2020; see section 1 - Introduction). More widely, our conclusion is also consistent with the difficulties that have been found to replicate many published results across the social and biomedical sciences, including psychology and neuroscience (Button et al., 2013; Ioannidis, 2005; Open Science Collaboration, 2015; Simmons et al., 2011).

Importantly, our findings should not be interpreted as speaking against the theory of embodiment in language. As stated by Simonsohn et al. (2014a), *p*-curve analysis evaluates the reliability of a set of results (whatever those findings are), not the theory they are supposed to be assessing. The tenet that the motor areas of the brain are functionally involved in action language comprehension may be right or wrong, but the present work cannot provide a conclusive answer to this question. However, what is clear is that, in order to evaluate a scientific question, we first need to have sound and reliable evidence. Present results do suggest that currently available neurostimulation studies of embodied language comprehension do not provide clear evidential value, contradicting the impression that a reader may obtain from the published evidence. Consequently, we encourage researchers to continue running TMS and tDCS studies for testing the predictions of the embodied view, but to do so keeping in mind the methodological recommendations described above, such as

preregistration, direct replication, statistical soundness, and the use of well-powered designs. Together with other high-quality studies from many different sources of evidence, including behavioral, brain imaging, psychophysiological, and patient studies, it will be possible to assess whether the meaning of action-related concepts is grounded on the motor system.

Declaration of interests

None.

Author contributions

Study conceptualization: PS and JS; Literature search: PS; Study selection: PS and JS; Contrast selection: PS and JS; Data analysis: PS; Writing - Original draft: PS; Writing - Review and Editing: JS; Funding acquisition: JS. 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.

Funding

This work was supported by the Project ref. PGC2018-096096-B-I00 from the Spanish Ministry of Science, Innovation, and Universities and the Project ref. PY20_00689 from the Andalusian Council and the European Regional Development Fund, both to JS, as well as by a FPU predoctoral grant (ref. FPU20/01946) to PS.

Acknowledgments

We are deeply grateful to the anonymous reviewers for providing us with valuable comments that have greatly improved the earlier version of this manuscript. We also thank Nikola

Vukovic, Yury Shtyrov, Karim Johari, Nicholas Riccardi, Svetlana Malyutina, Mirage Modi, Rutvik H. Desai, Valentina Niccolai, Anne Klepp, Peter Indefrey, Alfons Schnitzler, and Katja Biermann-Ruben for kindly discussing with us the statistical details of their studies. Finally, we thank the University of Granada for funding the Open Access Article charge of this publication.

References

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Curr. Biol., 16*(18), 1818-1823. https://doi.org/10.1016/j.cub.2006.07.060

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol., 59*, 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bestmann, S., & Krakauer, J. W. (2015). The uses and interpretations of the motor-evoked potential for understanding behaviour. *Exp. Brain Res.*, *233*(3), 679-689. https://doi.org/10.1007/s00221-014-4183-7

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends Cogn. Sci.*, *15*(11), 527-536. https://doi.org/10.1016/j.tics.2011.10.001

*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

Boulenger, V., Roy, A. C., Paulignan, Y., Deprez, V., Jeannerod, M., & Nazir, T. A. (2006). Cross-talk between language processes and overt motor behavior in the first 200 msec of processing. *J. Cogn. Neurosci., 18*(10), 1607-1615. https://doi.org/10.1162/jocn.2006.18.10.1607

*Branscheidt, M., Hoppe, J., Freundlieb, N., Zwitserlood, P., & Liuzzi, G. (2017). tDCS over the motor cortex shows differential effects on action and object words in associative word learning in healthy aging. *Front. Aging Neurosci.*, *9*:137. https://doi.org/10.3389/fnagi.2017.00137

*Branscheidt, M., Hoppe, J., Zwitserlood, P., & Liuzzi, G. (2018). tDCS over the motor cortex improves lexical retrieval of action words in poststroke aphasia. *J. Neurophysiol*, *119*(2), 621-630. https://doi.org/10.1152/jn.00285.2017

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

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. *Cognit. Brain Res., 24*(3), 355-363. https://doi.org/10.1016/j.cogbrainres.2005.02.020

*Bundt, C., Bardi, L., Abrahamse, E. L., Brass, M., & Notebaert, W. (2015). It wasn't me! Motor activation from irrelevant spatial information in the absence of a response. *Front. Hum. Neurosci.*, *9*:539. https://doi.org/10.3389/fnhum.2015.00539

Burns, E. J., Arnold, T., & Bukach, C. M. (2019). *P*-curving the fusiform face area: Meta-analyses support the expertise hypothesis. *Neurosci. Biobehav. Rev.*, *104*, 209-221. https://doi.org/10.1016/j.neubiorev.2019.07.003

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. *Nat. Rev. Neurosci., 14*(5), 365-376. https://doi.org/10.1038/nrn3475

*Cacciari, C., Bolognini, N., Senna, I., Pellicciari, M. C., Miniussi, C., & Papagno, C. (2011). Literal, fictive and metaphorical motion sentences preserve the motion component of

the verb: a TMS study. *Brain Lang.*, *119*(3), 149-157. https://doi.org/10.1016/j.bandl.2011.05.004

*Candidi, M., Leone-Fernandez, B., Barber, H. A., Carreiras, M., & Aglioti, S. M. (2010a). Hands on the future: facilitation of cortico-spinal hand-representation when reading the future tense of hand-related action verbs. *Eur. J. Neurosci.*, *32*(4), 677-683. https://doi.org/10.1111/j.1460-9568.2010.07305.x

*Candidi, M., Vicario, C. M., Abreu, A. M., & Aglioti, S. M. (2010b). Competing mechanisms for mapping action-related categorical knowledge and observed actions. *Cereb. Cortex, 20*(12), 2832-2841. https://doi.org/10.1093/cercor/bhq033

*Cattaneo, Z., Devlin, J. T., Salvini, F., Vecchi, T., & Silvanto, J. (2010). The causal role of category-specific neuronal representations in the left ventral premotor cortex (PMv) in semantic processing. *NeuroImage, 49*(3), 2728-2734. https://doi.org/10.1016/j.neuroimage.2009.10.048

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nat. Hum. Behav.*, *6*(1), 29-42. https://doi.org/10.1038/s41562-021-01193-7

Chatterjee, A. (2010). Disembodying cognition. *Lang. Cogn.*, *2*(1), 79-116. https://doi.org/10.1515/langcog.2010.004

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Routledge Academic.

*Courson, M., Macoir, J., & Tremblay, P. (2017). Role of medial premotor areas in action language processing in relation to motor skills. *Cortex*, *95*, 77-91. https://doi.org/10.1016/j.cortex.2017.08.002

De Marco, D., De Stefani, E., Bernini, D., & Gentilucci, M. (2018). The effect of motor context on semantic processing: a TMS study. *Neuropsychologia, 114*, 243-250. https://doi.org/10.1016/j.neuropsychologia.2018.05.003

de Vega, M., Moreno, V., & Castillo, D. (2013). The comprehension of action-related sentences may cause interference rather than facilitation on matching actions. *Psychol. Res., 77*(1), 20-30. https://doi.org/10.1007/s00426-011-0356-1

Dupont, W., Lebon, F., Papaxanthis, C., & Madden-Lombardi, C. (2020). The motor cortex wants the full story: The influence of sentence context on corticospinal excitability in action language processing. *HAL.* https://hal.archives-ouvertes.fr/hal-03053124.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol., 61*(6), 825-850. https://doi.org/10.1080/17470210701623605

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychon. Bull. Rev.*, *19*(2), 151-156. https://doi.org/10.3758/s13423-012-0227-9

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.*, *60*(4), 328-331. https://doi.org/10.1198/000313006X152649

*Gianelli, C., & Dalla Volta, R. (2015). Does listening to action-related sentences modulate the activity of the motor system? Replication of a combined TMS and behavioral study. *Front. Psychol., 5*:1511. https://doi.org/10.3389/fpsyg.2014.01511

*Gianelli, C., Kühne, K., Presti, S. L., Mencaraglia, S., & Dalla Volta, R. (2020). Action processing in the motor system: Transcranial Magnetic Stimulation (TMS) evidence of shared mechanisms in the visual and linguistic modalities. *Brain Cogn.*, *139*, 105510. https://doi.org/10.1016/j.bandc.2019.105510

*Gijssels, T., Ivry, R. B., & Casasanto, D. (2018). tDCS to premotor cortex changes action verb understanding: Complementary effects of inhibitory and excitatory stimulation. *Sci. Rep., 8*(1), 1-7. https://doi.org/10.1038/s41598-018-29600-6

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev., 9*(3), 558-565. https://doi.org/10.3758/BF03196313

*Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *Q. J. Exp. Psychol., 61*(6), 905-919. https://doi.org/10.1080/17470210701625550

*Gough, P. M., Campione, G. C., & Buccino, G. (2013). Fine tuned modulation of the motor system by adjectives expressing positive and negative properties. *Brain Lang.*, *125*(1), 54-59. https://doi.org/10.1016/j.bandl.2013.01.012

*Gough, P. M., Riggio, L., Chersi, F., Sato, M., Fogassi, L., & Buccino, G. (2012). Nouns referring to tools and natural objects differentially modulate the motor system. *Neuropsychologia, 50*(1), 19-25. https://doi.org/10.1016/j.neuropsychologia.2011.10.017

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*(2), 301-307. https://doi.org/10.1016/S0896-6273(03)00838-9

Hayek, D., Flöel, A., & Antonenko, D. (2018). Role of sensorimotor cortex in gestural-verbal integration. *Front. Hum. Neurosci., 12*:482. https://doi.org/10.3389/fnhum.2018.00482

*Innocenti, A., De Stefani, E., Sestito, M., & Gentilucci, M. (2014). Understanding of action-related and abstract verbs in comparison: a behavioral and TMS study. *Cogn. Process., 15*(1), 85-92. https://doi.org/10.1007/s10339-013-0583-z

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med., 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *J. Math. Psychol.*, *57*(5), 184-187. https://doi.org/10.1016/j.jmp.2013.03.002

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

*Johari, K., Riccardi, N., Malyutina, S., Modi, M., & Desai, R. H. (2021). HD-tDCS over motor cortex facilitates figurative and literal action sentence processing. *Neuropsychologia*, *159*, 107955. https://doi.org/10.1016/j.neuropsychologia.2021.107955

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Pers. Soc. Psychol. Rev., 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex, 48*(7), 805-825. https://doi.org/10.1016/j.cortex.2011.04.006

*Kuipers, J. R., van Koningsbruggen, M., & Thierry, G. (2013). Semantic priming in the motor cortex: evidence from combined repetitive transcranial magnetic stimulation and event-related potential. *Neuroreport*, *24*(12), 646-651. https://doi.org/10.1097/WNR.0b013e3283631467

*Labruna, L., Fernández-del-Olmo, M., Landau, A., Duqué, J., & Ivry, R. B. (2011). Modulation of the motor system during visual and auditory language processing. *Exp. Brain Res., 211*(2), 243-250. https://doi.org/10.1007/s00221-011-2678-z

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Jpn. Psychol. Rev.*, *62*(3), 221-230. https://doi.org/10.24602/sjpr.62.3_221

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

*Liuzza, M. T., Candidi, M., & Aglioti, S. M. (2011). Do not resonate with actions: sentence polarity modulates cortico-spinal excitability during action-related sentence reading. *PloS One*, *6*(2), e16855. https://doi.org/10.1371/journal.pone.0016855

*Liuzzi, G., Freundlieb, N., Ridder, V., Hoppe, J., Heise, K., Zimerman, M., ... & Hummel, F. C. (2010). The involvement of the left motor cortex in learning of a novel action word lexicon. *Curr. Biol., 20*(19), 1745-1751. https://doi.org/10.1016/j.cub.2010.08.034

*Lo Gerfo, E., Oliveri, M., Torriero, S., Salerno, S., Koch, G., & Caltagirone, C. (2008). The influence of rTMS over prefrontal and motor areas in a morphological task: grammatical vs. semantic effects. *Neuropsychologia*, *46*(2), 764-770. https://doi.org/10.1016/j.neuropsychologia.2007.10.012

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol.-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. *Adv. Methods Pract. Psychol. Sci.*, *3*(2), 185-199. https://doi.org/10.1177/2515245920902370

Meister, I. G., Wu, A. D., Deblieck, C., & Iacoboni, M. (2012). Early semantic and phonological effects on temporal-and muscle-specific motor resonance. *Eur. J. Neurosci., 36*(3), 2391-2399. https://doi.org/10.1111/j.1460-9568.2012.08134.x

Meteyard, L., Rodríguez-Cuadrado, S., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), 788-804. https://doi.org/10.1016/j.cortex.2010.11.002

*Monaco, E., Jost, L. B., Lancheros, M., Harquel, S., Schmidlin, E., & Annoni, J. M. (2021). First and Second Language at Hand: A Chronometric Transcranial-Magnetic Stimulation Study on Semantic and Motor Resonance. *J. Cogn. Neurosci.*, *33*(8), 1563-1580. https://doi.org/10.1162/jocn_a_01736

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., Ibáñez, A., García, A., Gianelli, C., Jones, J. L., Madden, J., Alifano, F., Bergen, B., Bloxsom, N. G., Bub, D. N., Cai, Z. C., Chartier, C. R., Chatterjee, A., Conwell, E., ... Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE*). Psychon. Bull. Rev*., *29*, 613-626. https://doi.org/10.3758/s13423-021-01927-8

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Murteira, A., Sowman, P. F., & Nickels, L. (2018). Does TMS disruption of the left primary motor cortex affect verb retrieval following exposure to pantomimed gestures? *Front. Neurosci., 12*:920. https://doi.org/10.3389/fnins.2018.00920

Navas, J. F., Verdejo-García, A., & Vadillo, M. A. (2021). The evidential value of research on cognitive training to change food-related biases and unhealthy eating behavior: A systematic review and *p*-curve analysis. *Obes. Rev.*, *22*(12), e13338. https://doi.org/10.1111/obr.13338

*Niccolai, V., Klepp, A., Indefrey, P., Schnitzler, A., & Biermann-Ruben, K. (2017). Semantic discrimination impacts tDCS modulation of verb processing. *Sci. Rep.*, *7*(1), 1-11. https://doi.org/10.1038/s41598-017-17326-w

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.*, *14*(9), 1105-1107. https://doi.org/10.1038/nn.2886

Nitsche, M. A., & Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *J. Physiol*, *527*(3), 633-639. https://doi.org/10.1111/j.1469-7793.2000.t01-1-00633.x

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline-Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). Replicability, Robustness, and Reproducibility in Psychological Science. *Annu. Rev. Psychol., 73,* 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods*, *48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

*Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A., & Pascual-Leone, A. (2004). All talk and no action: a transcranial magnetic stimulation study of motor cortex activation during action word production. *J. Cogn. Neurosci., 16*(3), 374-381. https://doi.org/10.1162/089892904322926719

*Onmyoji, Y., Kubota, S., Hirano, M., Tanaka, M., Morishita, T., Uehara, K., & Funase, K. (2015). Excitability changes in the left primary motor cortex innervating the hand muscles induced during speech about hand or leg movements. *Neurosci. Lett.*, *594*, 46-50. https://doi.org/10.1016/j.neulet.2015.03.052

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Ostarek, M., & Bottini, R. (2021). Towards strong inference in research on embodiment - Possibilities and limitations of causal paradigms. *J. Cogn, 4*(1), 5. http://doi.org/10.5334/joc.139

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Curr. Dir. Psychol. Sci., 28*(6), 593-599. https://doi.org/10.1177/0963721419866441

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Mother, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71. https://doi.org/10.1136/bmj.n71

*Papeo, L., Hochmann, J. R., & Battelli, L. (2016). The default computation of negated meanings. *J. Cogn. Neurosci.*, *28*(12), 1980-1986. https://doi.org/10.1162/jocn_a_01016

Papeo, L., Lingnau, A., Agosta, S., Pascual-Leone, A., Battelli, L., & Caramazza, A. (2015). The origin of word-related motor activity. *Cereb. Cortex, 25*(6), 1668-1675. https://doi.org/10.1093/cercor/bht423

Papeo, L., Pascual-Leone, A., & Caramazza, A. (2013). Disrupting the brain to validate hypotheses on the neurobiology of language. *Front. Hum. Neurosci., 7*:148. https://doi.org/10.3389/fnhum.2013.00148

*Papeo, L., Vallesi, A., Isaja, A., & Rumiati, R. I. (2009). Effects of TMS on different stages of motor and non-motor verb processing in the primary motor cortex. *PloS One, 4*(2), e4508. https://doi.org/10.1371/journal.pone.0004508

Papesh, M. H. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *J. Exp. Psychol.-Gen., 144*(6), e116. https://doi.org/10.1037/xge0000125

*Papitto, G., Lugli, L., Borghi, A. M., Pellicano, A., & Binkofski, F. (2021). Embodied negation and levels of concreteness: A TMS Study on German and Italian language processing. *Brain Res.*, *1767*, 147523. https://doi.org/10.1016/j.brainres.2021.147523

Pessoa, L. (2014). Understanding brain networks and brain organization. *Phys. Life Rev., 11*(3), 400-435. https://doi.org/10.1016/j.plrev.2014.03.005

Polanía, R., Nitsche, M.A. & Ruff, C.C. (2018). Studying and modifying brain function with non-invasive brain stimulation. *Nat. Neurosci., 21*, 174–187. https://doi.org/10.1038/s41593-017-0054-4

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nat. Rev. Neurosci.*, *6*(7), 576–582. https://doi.org/10.1038/nrn1706

*Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *Eur. J. Neurosci., 21*(3), 793-797. https://doi.org/10.1111/j.1460-9568.2005.03900.x

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

*Reilly, M., Howerton, O., & Desai, R. H. (2019). Time-course of motor involvement in literal and metaphoric action sentence processing: A TMS study. *Front. Psychol*, *10*:371. https://doi.org/10.3389/fpsyg.2019.00371

*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

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

Saccone, E. J., Thomas, N. A., & Nicholls, M. E. R. (2021). One-handed motor activity does not interfere with naming lateralized pictures of tools. *J. Exp. Psychol.-Hum. Percept. Perform., 47*(4), 529–544. https://doi.org/10.1037/xhp0000863

Schomers, M. R., Kirilina, E., Weigand, A., Bajbouj, M., & Pulvermüller, F. (2015). Causal influence of articulatory motor cortex on comprehending single spoken words: TMS evidence. *Cereb. Cortex, 25*(10), 3894-3902. https://doi.org/10.1093/cercor/bhu274

*Scorolli, C., Jacquet, P. O., Binkofski, F., Nicoletti, R., Tessari, A., & Borghi, A. M. (2012). Abstract and concrete phrases processing differentially modulates cortico-spinal excitability. *Brain Res., 1488*, 60-71.

Shebani, Z., & Pulvermüller, F. (2018). Flexibility in language action interaction: the influence of movement type. *Front. Hum. Neurosci., 12*:252. https://doi.org/10.1016/j.brainres.2012.10.004

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci., 22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration is a Game Changer. But, Like Random Assignment, it is Neither Necessary Nor Sufficient for Credible Science. *J. Consum. Psychol.*, *31*(1), 177-180. https://doi.org/10.1002/jcpy.1207

Simmons, J. P., & Simonsohn, U. (2017). Power posing: *P*-curving the evidence. *Psychol. Sci.*, *28*(5), 687-693. https://doi.org/10.1177/0956797616658563

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *P*-curve: a key to the file-drawer. *J. Exp. Psychol.-Gen., 143*(2), 534-547. https://doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspect. Psychol. Sci., 9*(6), 666-681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *P*-curves: Making *P*-curve analysis more robust to errors, fraud, and ambitious *P*-hacking, a Reply to Ulrich and Miller (2015). *J. Exp. Psychol.-Gen., 144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

*Suárez-García, D., Birba, A., Zimerman, M., Diazgranados, J. A., Lopes da Cunha, P., Ibáñez, A., Grisales-Cárdenas, J. S., Cardona, J. F., & García, A. M. (2021). Rekindling action language: a neuromodulatory study on Parkinson's disease patients. *Brain Sci.*, *11*(7), 887. https://doi.org/10.3390/brainsci11070887

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., & Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci., 17*(2), 273-281. https://doi.org/10.1162/0898929053124965

Togato, G., Andras, F., Miralles, E., & Macizo, P. (2021). Motor processing modulates word comprehension. *Br. J. Psychol.*, *112*(4), 1028-1052. https://doi.org/10.1111/bjop.12507

*Tomasino, B., Fink, G. R., Sparing, R., Dafotakis, M., & Weiss, P. H. (2008). Action verbs and the primary motor cortex: a comparative TMS study of silent reading, frequency judgments, and motor imagery. *Neuropsychologia, 46*(7), 1915-1926. https://doi.org/10.1016/j.neuropsychologia.2008.01.015

*Tremblay, P., Sato, M., & Small, S. L. (2012). TMS-induced modulation of action sentence priming in the ventral premotor cortex. *Neuropsychologia*, *50*(2), 319-326. https://doi.org/10.1016/j.neuropsychologia.2011.12.002

Ulrich, R., & Miller, J. (2015). *p*-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *J. Exp. Psychol.-Gen., 144*(6), 1137–1145. https://doi.org/10.1037/xge0000086

Vadillo, M. A., Gold, N., & Osman, M. (2016a). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychol. Sci.*, *27*(9), 1207-1214. https://doi.org/10.1177/0956797616654911

Vadillo, M. A., Hardwicke, T. E., & Shanks, D. R. (2016b). Selection bias, vote counting, and money-priming effects: A comment on Rohrer, Pashler, and Harris (2015) and Vohs (2015). *J. Exp. Psychol.-Gen.*, *145*(5), 655–663. https://doi.org/10.1037/xge0000157

*Vicario, C. M., Candidi, M., & Aglioti, S. M. (2013). Cortico-spinal embodiment of newly acquired, action-related semantic associations. *Brain Stimul.*, *6*(6), 952-958. https://doi.org/10.1016/j.brs.2013.05.010

*Vicario, C. M., & Rumiati, R. I. (2012). tDCS of the primary motor cortex improves the detection of semantic dissonance. *Neurosci. Lett.*, *518*(2), 133-137. https://doi.org/10.1016/j.neulet.2012.04.070

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

*Vukovic, N., & Shtyrov, Y. (2019). Learning with the wave of the hand: Kinematic and TMS evidence of primary motor cortex role in category-specific encoding of word meaning. *NeuroImage, 202*, 116179. https://doi.org/10.1016/j.neuroimage.2019.116179

*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

Vukovic, N., Hansen, B., Lund, T. E., Jespersen, S., & Shtyrov, Y. (2021). Rapid microstructural plasticity in the cortical semantic network following a short language learning session. *PLoS Biol., 19*(6), e3001290. https://doi.org/10.1371/journal.pbio.3001290

Walsh, V., & Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nat. Rev. Neurosci.*, *1*(1), 73-80. https://doi.org/10.1038/35036239

*Willems, R. M., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding: evidence from theta-burst

transcranial magnetic stimulation. *Psychol. Sci., 22*(7), 849-854. https://doi.org/10.1177/0956797611412387

Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. (2010). A functional role for motor simulation in identifying tools. *Psychol. Sci., 21*(9), 1215-1219. https://doi.org/10.1177/0956797610378307

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). *Psychol. Sci.*, *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. *Behav. Brain Sci., 41*, E120. doi:10.1017/S0140525X17001972

Figures

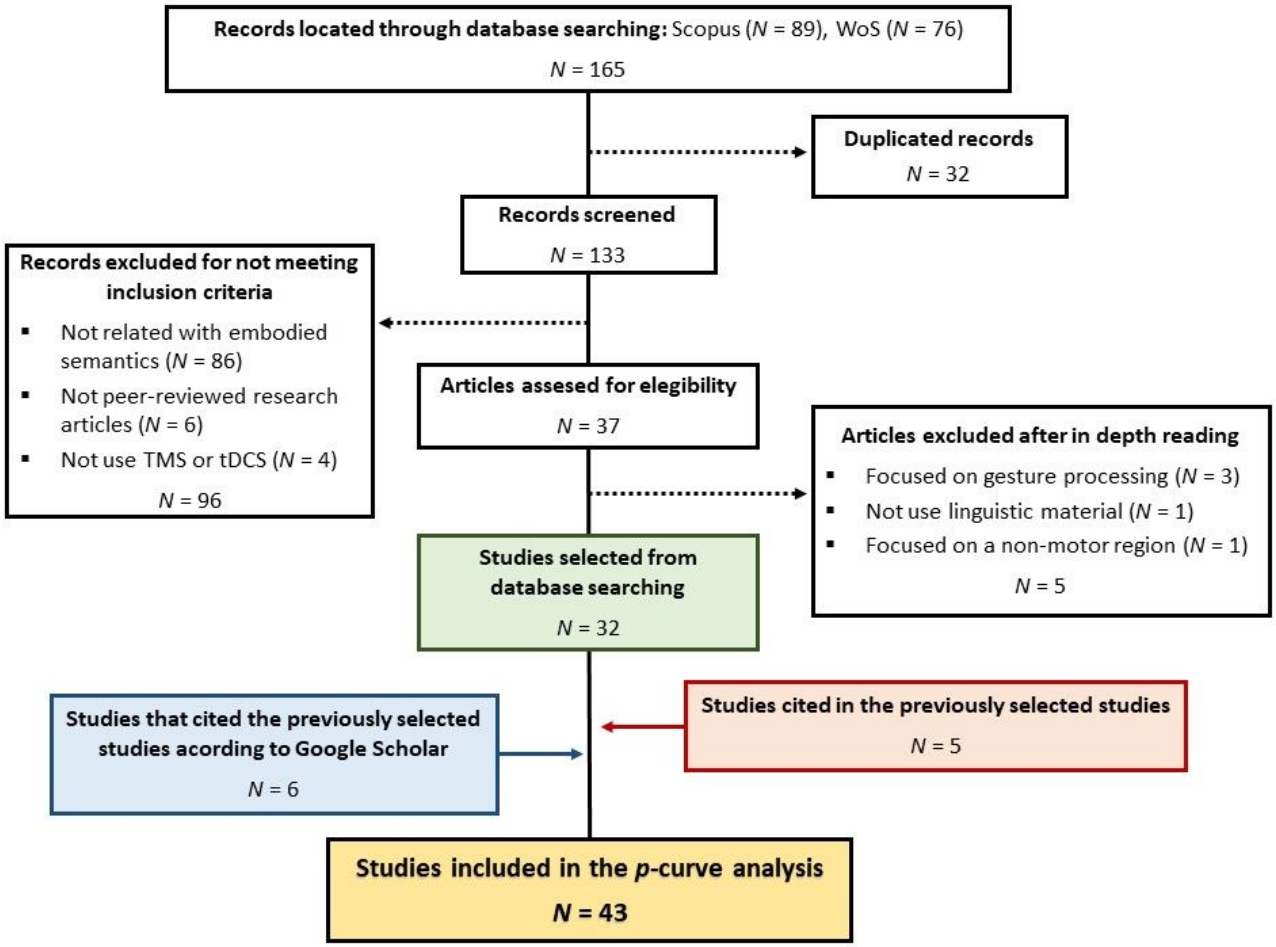

Figure 1. Flowchart of the literature search and the article selection process.

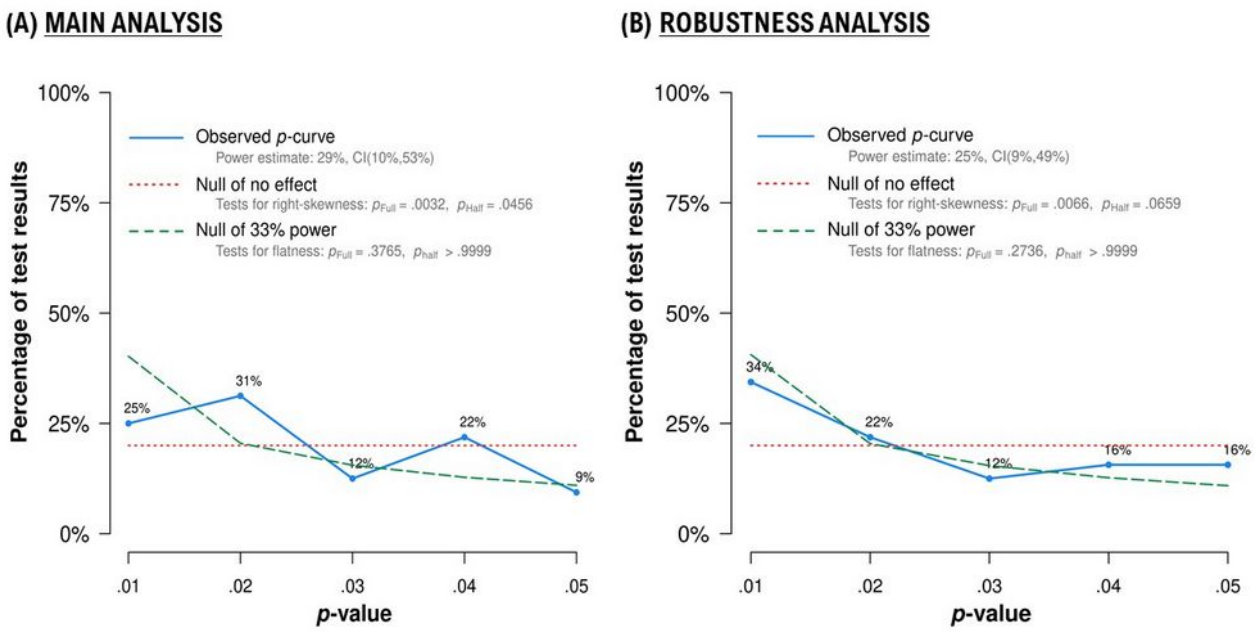

Figure 2. Distribution of the selected *p*-values for the main analysis (A) and the robustness analysis (B). The red dotted line represents the expected distribution if the studies explore null effects. The green striped line represents the expected distribution if the studies explore true effects but with a power of only 33%. The blue continuous line depicts the observed distribution of *p*-values.

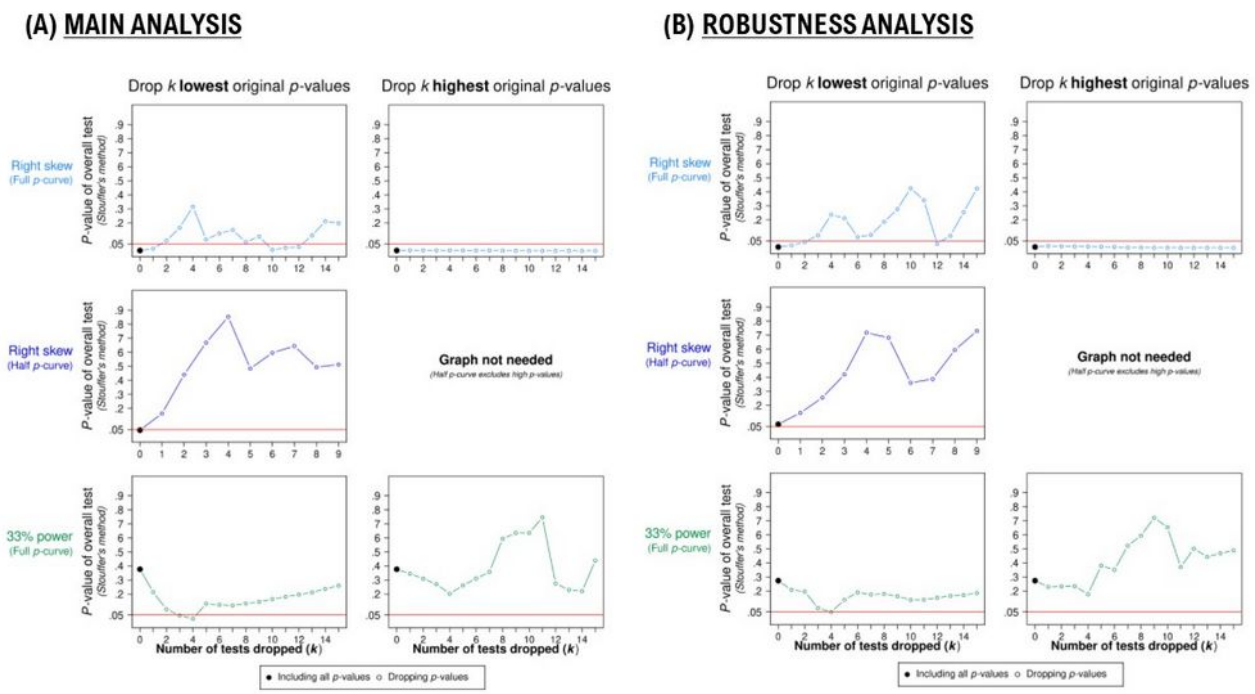

Figure 3. Cumulative meta-analysis for the main analysis (A) and the robustness analysis (B). Plots represent how the significance level of the right-skewness test for the full (top) and half *p*-curve (mid), and the flatness test (bottom) changes if we progressively exclude the most extreme *p*-values included in the analysis until reaching half of them. The red horizontal line depicts the conventional significance threshold of $p = 0.05$.

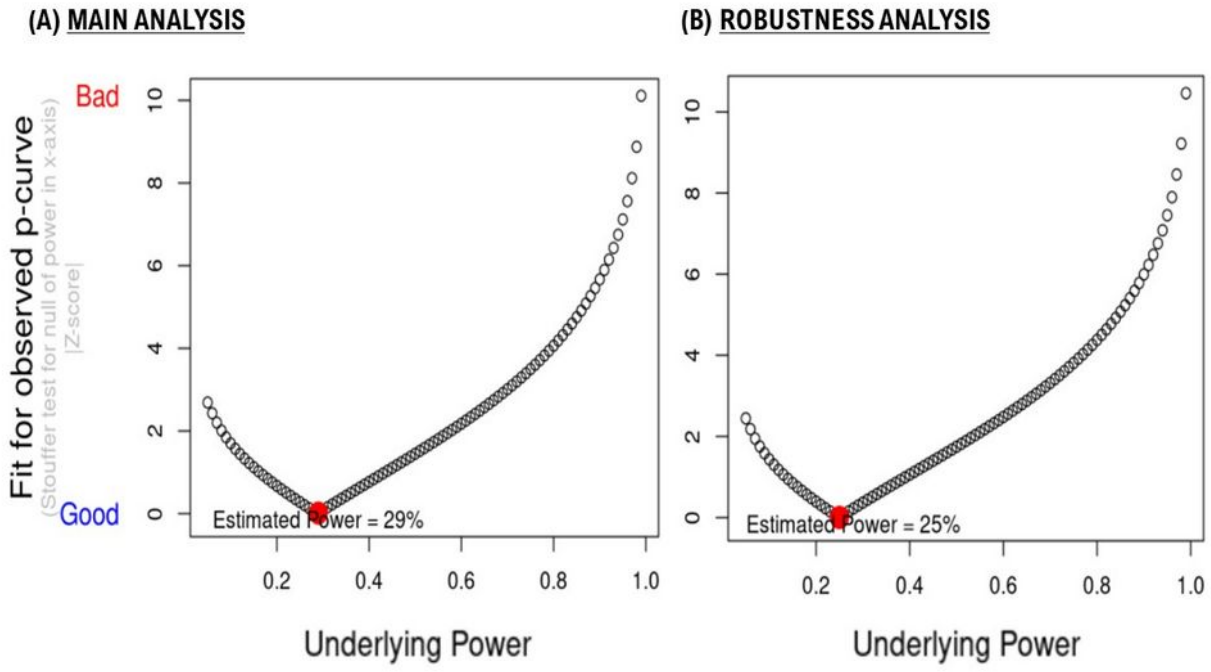

Figure 4. Estimation of the underlying power of the studies included in the main analysis (A) and the robustness analysis (B). Plots represent the degree of fit (vertical axis) between the present *p*-curve and the expected *p*-curves for each value ranging between 5% and 99% of power (horizontal axis). The resulting estimate (red point) is the value with the better fit.

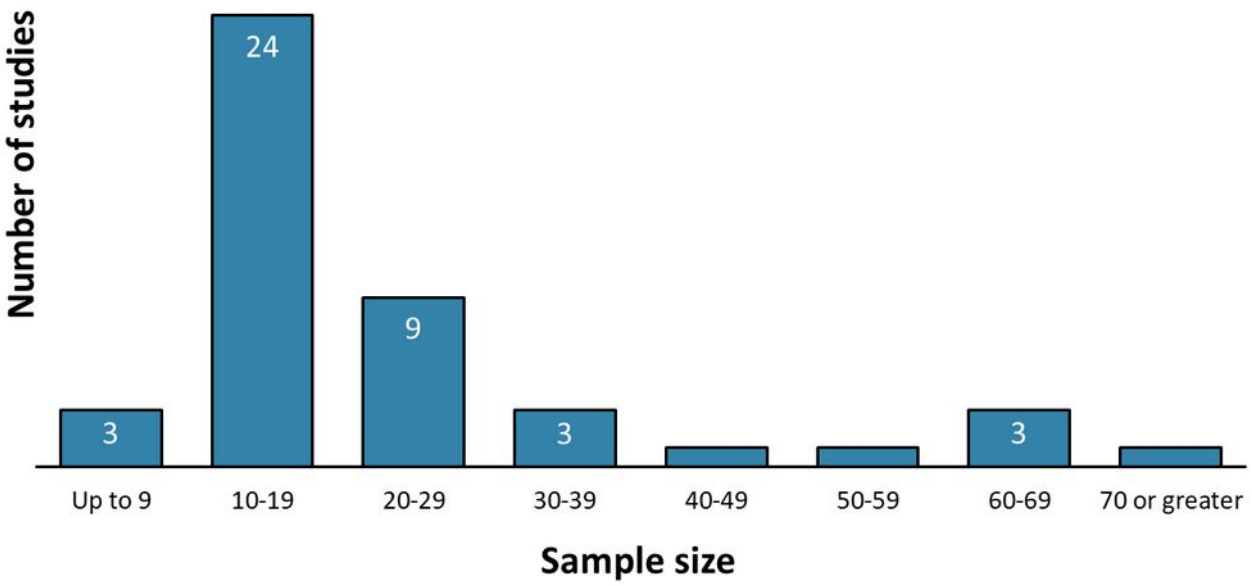

Figure 5. Sample size distribution across the studies included in the analyses.

Tables

Table 1. Main characteristics of the studies included in the *p*-curve analysis.

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
Birba et al. (2020)	*N* = 68	tDCS	M1	Text type (action-landen vs. neutral) x Information type (action-related vs. circumstantial) x Stimulation (anodal-M1 vs. anodal-VLPFC vs. sham-M1)	Answering comprehension questions about the texts	Performance in the text comprehension task	Supports
Branscheidt et al. (2017)	*N* = 18 (Healthy old participants)	tDCS	M1	Word type (action verb vs. object noun) x Stimulation (anodal vs. cathodal vs. sham)	Translate pseudowords into the participants' language	Percentage of novel action and object words correctly translated	Supports
Branscheidt et al. (2018)	*N* = 16 (Post-stroke aphasic participants)	tDCS	M1	Word type (action verbs vs. object nouns) x Stimulus type (real word vs. pseudoword) x Stimulation (anodal vs. sham)	Lexical decision	RT + ACC	Supports
Buccino et al. (2005) *Experiment 1*	*N* = 8	sp-TMS	M1	Sentence type (hand-related vs. foot-related vs. abstract) x Stimulation site (hand M1 vs. foot M1) x Muscle (opponens pollicis vs. first dorsal	Passive listening	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				interosseus for hand; tibialis anterior vs. gastrocnemius for foot)			
Bundt et al. (2015)	*N* = 22	sp-TMS	M1	Word ("right" vs. "left") x Trial (compatible vs. incompatible vs. neutral) x Stimulation site (right M1 vs. left M1) x TMS timing (250 vs. 320 vs. 500 vs. 640 ms)	Discriminate the color of a cross (color trials) and silently read words (word trials)	MEPs amplitude	Supports
Cacciari et al. (2011)	*N* = 9	sp-TMS	M1	Sentence type (literal motion vs. metaphorical motion vs. fictive motion vs. idiomatic vs. mental verbs) x Muscle (tibialis anterior vs. gastrocnemius)	Silent reading	MEPs amplitude	Supports
Candidi et al. (2010a)	*N* = 19	sp-TMS	M1	Verb type (hand-related vs. foot-related vs. sensory vs. abstract) x Verb tense (future vs. past) x Muscle (first dorsal interosseous vs. tibialis anterior)	Reading aloud verbs	MEPs amplitude	Supports
Candidi et al. (2010b) *Experiment 1*	*N* = 13	sp-TMS	M1	Stimulus type (surname vs. face) x Sport (football vs. tennis) x Limb (arm vs. leg) x Muscle (tibialis anterior vs. soleus for leg; extensor carpi radialis vs.	Discriminate between soccer players and tennis players	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				flexor carpi radialis for arm)			
Cattaneo et al. (2010)	*N* = 12	sp-TMS	vPMC	Word type (tool noun vs. animal noun) x Trial type (congruent vs. incongruent) x Stimulation (vPMC vs. dPMC vs. No TMS)	Discriminate between tool nouns and animal nouns	RT	Supports
Courson et al. (2017) *Study 2*	*N* = 16	rTMS	SMA	Sentence type (human action vs. non-human action) x Stimulation site (SMA vs. pre-SMA) x Stimulation (TMS vs. No TMS)	Determine if the content of a sentence was true or false	RT + ACC	Supports
Gianelli & Dalla Volta (2015) *Experiment 1*	*N* = 21 (Pre-registered)	sp-TMS	M1	Sentence type (hand-related vs. foot-related vs. abstract) x Stimulation site (hand M1 vs. foot M1)	Passive listening	MEPs amplitude	Supports
Gianelli et al. (2020) *Experiment 1*	*N* = 14	sp-TMS	M1	Modality (linguistic: pairs of object nouns and action verbs vs. visual: pairs of images representing objects and actions) x Muscle (abductor digiti minimi vs. first dorsal interosseous) x TMS timing (baseline vs. object vs. 150 vs. 350 vs.	Answer questions related to the presented stimuli	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				500 ms) Language: L1			
Gianelli et al. (2020) *Experiment 2*	*N* = 14	sp-TMS	M1	Modality (linguistic: pairs of object nouns and action verbs vs. visual: pairs of images representing objects and actions) x Muscle (abductor digiti minimi vs. first dorsal interosseous) x TMS timing (baseline vs. object vs. 150 vs. 350 vs. 500 ms) Language: L2	Answer questions related to the presented stimuli	MEPs amplitude	Supports
Gijssels et al. (2018)	*N* = 73	tDCS	PMC	Word type (action verb vs. abstract verb) x Stimulation (anodal vs. cathodal) x Response hand (right vs. left)	Lexical decision	RT + ACC	Supports
Glenberg et al. (2008) *Experiment 2*	*N* = 11	sp-TMS	M1	Verb type (concrete vs. abstract) x Sentence type (transfer vs. no-transfer) x TMS timing (at the end of the verb vs. at the end of the sentence)	Discriminate between sensible and non-sensible phrases	MEPs amplitude	Supports
Gough et al. (2012)	*N* = 15	sp-TMS	M1	Word type (graspable noun vs. non-graspable	Discriminate between	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				noun) x Noun type (natural vs. artificial)	objects and non-objects nouns		
Gough et al. (2013)	*N* = 14	sp-TMS	M1	Adjective type (denoting a positive vs. a negative property for interaction) x Muscle (first dorsal interosseus vs. extensor communis digitorum)	Discriminate if a letter was present in the later presented word	MEPs amplitude	Supports
Innocenti et al. (2014) *TMS Experiment*	*N* = 13	sp-TMS	M1	Word type (action verb vs. abstract verb) x Block (first vs. second)	Discriminate between action verbs and abstract verbs	MEPs amplitude	Supports
Johari et al. (2021)	*N* = 23	HD-tDCS	M1	Sentence type (literal action vs. idiomatic action vs. metaphoric action vs. visual) x Stimulation (cathodal vs. sham)	Discriminate between sensible and non-sensible phrases	RT	Supports
Kuipers et al. (2013)	*N* = 12	rTMS	M1	Word type (hand-related verb vs. mouth-related verb) x Stimulation (active TMS vs. sham) x Electrode (Cz vs. C2 vs. CPz vs. CP2)	Read pairs of verbs silently	N400 amplitude	Supports
Labruna et al. (2011)	*N* = 19	sp-TMS	M1	Word type (action-related verb vs. non-action word) x Modality (visual vs. auditory) x TMS timing	Silent reading + Passive listening	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				(150 vs. 300 ms)			
Liuzza et al. (2011)	*N* = 14	pp-TMS	M1	Sentence type (action-related vs. abstract) x Sentence polarity (positive vs. negative)	Answering questions concerning the last read sentence	MEPs amplitude	Supports
Liuzzi et al. (2010)	*N* = 63	tDCS	M1	Word type (pseudoword associated with a body-related action vs. associated with an object) x Stimulation (anodal vs. cathodal vs. sham) x Stimulation site (M1 vs. DLPFC)	Translate pseudowords into the participants' language	Percentage of novel action words correctly translated	Supports
Lo Gerfo et al. (2008) *Experiment 2*	*N* = 15	rTMS	M1	Grammatical class (verb vs. noun) x Semantic class (action-related vs. abstract) x Stimulation (TMS vs. No TMS)	Produce singular/plural forms of nouns + Conjugate verbs	RT	Supports
Monaco et al. (2021)	*N* = 34	sp-TMS	M1	Word type (action verb vs. non-action verb) x Language (L1 vs. L2) x TMS timing (125 vs. 275 vs. 350 vs. 500 ms)	Decide if the presented verbs describe a physical or a mental action	RT + MEPs amplitude	Supports
Niccolai et al. (2017)	*N* = 20	tDCS	M1	Word type (hand-related verb vs. foot-related verb) x Stimulation (anodal vs. cathodal vs. sham) x Response effector (hand vs. foot) x Semantic	Discriminate between concrete and abstract verbs using either the hand or the	RT + ACC	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				discrimination performance (high vs. low)	foot, depending on the geometric shape of a prime		
Oliveri et al. (2004)	*N* = 8	sp-TMS + pp-TMS	M1	Grammatical class (verb vs. noun) x Semantic class (action vs. non-action) x Stimulation (single-pulse vs. paired-pulse ISI 1 ms vs. paired-pulse ISI 10 ms)	Produce singular/plural forms of nouns + Conjugate verbs	MEPs amplitude	Supports
Onmyoji et al. (2015) *Experiment 1*	*N* = 18	sp-TMS	M1	Sentence type (hand movement vs. foot movement vs. no movement) x Reading (aloud vs. silent) x TMS timing (1000 vs. 2000 ms)	Silent reading + Reading aloud	MEPs amplitude	Against
Papeo et al. (2009) *Experiment 1*	*N* = 11	sp-TMS	M1	Word type (hand action verb vs. non-hand action verb vs. non-action verb) x Task (semantic vs. syllabic) TMS timing: 170 ms	Discriminate between action and non-action verbs (semantic task) + Indicate the number of syllables of the verbs (syllabic task)	MEPs amplitude + RT + ACC	Against
Papeo et al.	*N* = 14	sp-TMS	M1	Word type (hand action	Discriminate	MEPs	Against

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
(2009) *Experiment 2*				verb vs. non-hand action verb vs. non-action verb) x Task (semantic vs. syllabic) TMS timing: 350 ms	between action and non-action verbs (semantic task) + Indicate the number of syllables of the verbs (syllabic task)	amplitude + RT + ACC	
Papeo et al. (2009) *Experiment 3*	*N* = 11	sp-TMS	M1	Word type (hand action verb vs. non-hand action verb vs. non-action verb) x Task (semantic vs. syllabic) TMS timing: 500 ms	Discriminate between action and non-action verbs (semantic task) + Indicate the number of syllables of the verbs (syllabic task)	MEPs amplitude + RT + ACC	Against
Papeo et al. (2016) *Experiment 1*	*N* = 18	sp-TMS	M1	Word type (action-related verb vs. state-related verb) x Polarity (positive vs. negative) x TMS timing (250 vs. 400 vs. 550 ms)	Recognize the already presented verbs	MEPs amplitude	Supports
Papeo et al. (2016) *Experiment 2*	*N* = 14	sp-TMS	M1	Word type (action-related verb vs. state-related verb) x Polarity (positive vs. negative)	Recognize the already presented verbs	EMG: cortical silent period (CSP) duration	Supports
Papitto et al.	*N* = 42	sp-TMS	M1	Concreteness (abstract	Silent reading	MEPs	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
(2021)				verb + abstract noun vs. abstract verb + concrete noun vs. concrete verb + concrete noun) x Polarity (positive vs. negative) x TMS timing (at verb vs. at noun vs. at adverb) x Language (Italian vs. German)		amplitude	
Pulvermüller et al. (2005)	*N* = 12	sp-TMS	M1	Word type (hand-related verb vs. foot-related verb) x Stimulation site (hand M1 vs. foot M1)	Lexical decision	Latency of the EMG recordings (RT)	Supports
Reilly et al. (2019)	*N* = 33	sp-TMS	M1	Sentence type (literal action vs. metaphoric action vs. abstract) x Stimulation site (M1 vs. occipital pole) x TMS timing (150 vs. 300 vs. 450 ms)	Discriminate between sensible and non-sensible phrases	RT + MEPs amplitude	Supports
Repetto et al. (2013)	*N* = 20	rTMS	M1	Word type (action verb vs. abstract verb) x Stimulation site (right M1 vs. left M1)	Discriminate between concrete and abstract verbs	RT	Supports
Scorolli et al. (2012)	*N* = 16	sp-TMS	M1	Verb type (concrete vs. abstract) x Noun type (concrete vs. abstract) x Sentence type (sensible vs. non-sensible) x Stimulation (active TMS	Discriminate between sensible and non-sensible phrases	MEPs amplitude + RT + ACC	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				vs. sham) x TMS timing (at verb vs. at noun)			
Suárez-García et al. (2021)	*N* = 22 (Parkinson's Disease patients)	tDCS	M1	Word type (action verb vs. object noun) x Stimulation (anodal vs. sham) x Time point (before vs. after stimulation)	Word-picture association	ACC + RT	Supports
Tomasino et al. (2008)	*N* = 20	sp-TMS	M1	Task (silent reading vs. frequency judgement vs. motor imagery) x Stimulation site (M1 vs. vertex) x Stimulation timing (150 vs. 300 vs. 450 ms) Stimuli: action verbs	Silent reading + Frequency judgment + Motor imagery	Participants' judgements in the task + RT	Against
Tremblay et al. (2012)	*N* = 16	rTMS	vPMC	Sentence type (manual action vs. manipulable noun vs. orofacial vs. non-manipulable noun) x Stimulation (active vs. sham TMS)	Decide if a target word was semantically congruent with a phrase or not	RT + ACC	Supports
Vicario & Rumiati (2012)	*N* = 36	tDCS	M1	Sentence type (motor vs. non-motor) x Stimulation (anodal vs. cathodal vs. sham) x Trial (matching vs. mismatching)	Sentence-picture association	RT	Supports
Vicario et al. (2013)	*N* = 14	sp-TMS	M1	Word type (soccer player name vs. tennis player	Discriminate between	MEPs amplitude	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
				name vs. actor name) x Muscle (extensor carpi radialis vs. tibialis anterioris) x Time after learning (0 vs. 24 vs. 72 h)	soccer-related names and tennis-related names		
Vitale et al. (2021)	*N* = 50	tDCS	M1	Sentence type (action vs. attentional) x Stimulation (anodal-active vs. anodal-sham vs. cathodal-active vs. cathodal-sham)	Memorize and recall phrases	ACC + MEPs amplitude	Supports
Vukovic et al. (2017)	*N* = 28	rTMS	M1	Word type (concrete verb vs. abstract verb in the concreteness judgement task; or word vs. pseudoword in the lexical decision task) x Stimulation site (right M1 vs. left M1 vs. No TMS) x Task (concreteness judgement vs. lexical decision)	Concreteness judgement + Lexical decision	RT	Supports
Vukovic & Shtyrov (2019)	*N* = 68	cTBS	M1	Word type (verb vs. noun) x Stimulation (M1-TMS vs. SPL-TMS vs. M1-sham) x Word novelty (old vs. new) x Block (1-7)	Word learning + Lexical decision	ACC + RT + Movement complexity	Supports
Willems et al. (2011)	*N* = 20	cTBS	PMC	Word type (manual verb vs. non-manual verb) x	Lexical decision	RT	Supports

Study	Sample size	Stimulation type	Stimulation site	Experimental design	Experimental task	Dependent measure	Conclusions regarding the embodiment hypothesis
Main Experiment				Stimulation site (right PMC vs. left PMC)			

Abbreviations: tDCS = Transcranial Direct Current Stimulation; sp-TMS = Single-Pulse Transcranial Magnetic Stimulation; pp-TMS: Paired-Pulse Transcranial Magnetic Stimulation; rTMS = Repetitive Transcranial Magnetic Stimulation; cTBS = Continuous Theta-Burst Stimulation; M1 = Primary Motor Cortex; PMC = Premotor Cortex; vPMC = Ventral Premotor Cortex; dPMC = Dorsal Premotor Cortex; SMA = Supplementary Motor Area; VLPFC = Ventrolateral Prefrontal Cortex; DLPFC = Dorsolateral Prefrontal Cortex; SPL = Superior Parietal Lobule; RT = Reaction Time; ACC = Accuracy; MEPs = Motor Evoked Potentials; EMG = Electromyography.

Note: The column "Conclusions regarding the embodiment hypothesis" includes the conclusions manifested by the authors of the paper, regardless of whether these studies provided significant *p*-values for the present analyses or not (for detailed information on the contrast selection process, see section 2.4 - Contrast Selection and Supplementary Material, Appendix 2 - Disclosure table).