Title:

Does the motor system functionally contribute to keeping words in working memory? A pre-registered replication of Shebani and Pulvermüller (2013, Cortex)

Authors:

Guillermo Montero-Melis [a b c], Jeroen van Paridon [a], Markus Ostarek [a], Emanuel Bylund [c d]

[a] Max Planck Institute for Psycholinguistics

[b] Department of Linguistics, Stockholm University

[c] Centre for Research on Bilingualism, Stockholm University

[d] Department of General Linguistics, Stellenbosch University

Corresponding author:

Guillermo Montero-Melis
Max Planck Institute for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen, The Netherlands
Guillermo.MonteroMelis@mpi.nl
+31 24 3521322

Author note:

In accordance with the Peer Reviewers' Openness Initiative (Morey et al., 2016), all materials and scripts associated with this manuscript are available at https://osf.io/ktsfw/?view_only=63e3071ba35641a0ba11785324e427e3 (see list of appendices at the end of the manuscript).

ABSTRACT

Increasing evidence implicates the sensorimotor systems with high-level cognition, but the extent to which these systems play a functional role remains debated. Using an elegant design, Shebani and Pulvermüller (2013) reported that carrying out a demanding rhythmic task with the hands led to selective impairment of working memory for hand-related words (e.g., clap), while carrying out the same task with the feet led to selective memory impairment for foot-related words (e.g., kick). This striking double dissociation constitutes strong evidence for an embodied account of working memory, a system that has received scarce attention in the embodiment literature. However, the original study was likely underpowered and used suboptimal statistical analyses, raising questions about the robustness of the results. In fact, our reanalysis with improved statistical tools reproduced the critical interaction but yielded a substantially reduced effect size. In order to re-examine this seminal finding, we here attempt a fully pre-registered replication of the original study, following a sequential design with a maximal participant sample size over six times that of the original (96 vs. 15). We will analyse the results with Bayesian generalized mixed models and use Bayes factors to decide if the evidence supports the effect predicted by theories of embodied cognition.

Keywords:

Embodiment, working memory, semantics, action words, replication, registered report

1. Introduction

What is the nature of the system underlying high-level cognitive functions in the human brain? The traditional view from cognitive science is that high-level cognition is achieved by an amodal symbol system that is separated from the sensory and motor systems (Fodor, 1975; Newell, 1980; Pylyshyn, 1980). An opposing view that has gained scientific support in the last two decades claims that cognition is *embodied*, ascribing a central role to sensorimotor systems in various high-level cognitive processes, including access to meaning during language processing (Aziz-Zadeh & Damasio, 2008; Barsalou, 2008; Gallese & Lakoff, 2005; Pulvermüller, 2005; Pulvermüller & Fadiga, 2010). An interesting initial finding supporting embodied meaning representations is that action word semantics have a correlate in somatotopic activation of the motor cortex. For example, when people passively read words that denote actions carried out with different body parts – such as *lick* (tongue), *pick* (arm) or *kick* (leg) – similar parts of their motor and premotor cortex are activated as when they actually move the corresponding body parts (Hauk, Johnsrude, & Pulvermüller, 2004; Pulvermüller, Kherif, Hauk, Mohr, & Nimmo-Smith, 2009; Raposo, Moss, Stamatakis, & Tyler, 2009; Shtyrov, Butorina, Nikolaeva, & Stroganova, 2014; Tettamanti et al., 2005). However, such patterns of activation do not per se show that effector-specific motor processes are *causally* involved in processing the meaning of action words (Hickok, 2010; Mahon, 2015; Mahon & Caramazza, 2008).

A strong test of the functional relevance of the motor system for semantic processing comes from interference paradigms in healthy individuals. These paradigms typically have participants process action-related language while either disrupting cortical activity in motor areas with transcranial magnetic stimulation (e.g., Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005; Tomasino, Fink, Sparing, Dafotakis, & Weiss, 2008; Vukovic, Feurra, Shpektor, Myachykov, &

Shtyrov, 2017) or having them carry out a concurrent motor task (e.g., Boulenger et al., 2006; Yee, Chrysikou, Hoffman, & Thompson-Schill, 2013). A causal role can be inferred if activating parts of our motor system that map onto specific body parts (e.g., the arms) selectively interferes with processing of action words that refer to arm-related actions (e.g., clap), but not with words that relate to other body parts (e.g., kick) (see García & Ibáñez, 2016 for a comprehensive review of behavioural studies).

Interference is also a common method in studies on working memory, where interaction between a concurrent task (e.g., motor movements) and working memory performance provides evidence that both tasks are supported by the same function. Under the embodiment view that memory works in the service of action and perception, such interactions are expected (Barsalou, 1999; Glenberg, 1997). More generally, a central debate in this literature concerns the type of representations working memory operates on: Under the classical multi-component view, working memory acts as an autonomous buffer that operates independently of long-term memory and of the sensory and motor systems (Baddeley, 2003; Baddeley & Dale, 1966; Baddeley & Hitch, 1974). In contrast, recent state-based models do not posit separate components for long- and short-term representations, but instead assume that working memory consists in the allocation of attention to essentially the same internal representations as used in non-mnemonic settings (D'Esposito & Postle, 2015). This latter class of models starts from the premise that the same sensorimotor systems used to perceive information also contribute to the retention of that information in working memory (Awh & Jonides, 2001; Pasternak & Zaksas, 2003; Postle, Idzikowski, Sala, Logie, & Baddeley, 2006). Under the assumption that word meanings are (partly) constituted of sensorimotor representations, state-based models more naturally accommodate embodiment effects when verbal stimuli have to be kept in working memory, compared to models that posit a separate buffer.

Much of the previous evidence investigating whether motor simulations are involved in working memory has targeted the domain of object memory. These studies start from the central finding that motor affordances (such as the particular hand shape with which an object is grasped) are automatically activated during object perception even when they are task irrelevant (Tucker & Ellis, 1998, 2001). Support for a role of motor affordances in working memory comes from paradigms in which to-be-remembered objects are preceded by either a congruent or incongruent grasping movement: congruent pairs are better remembered than incongruent ones, suggesting that activating actions associated with the objects supports recall (Downing-Doucet & Guérard, 2014; see also Guérard, Guerrette, & Rowe, 2015; Lagacé & Guérard, 2015). These affordances also seem to play a role for the retention of words denoting objects (rather than pictures of objects). Dutriaux and colleagues recently showed that manipulable objects were better remembered with the hands free than when keeping the hands crossed behind the back, while this manipulation did not affect memory for non-manipulable objects; importantly, this effect persisted when words (instead of images) were shown (Dutriaux, Dahiez, & Gyselinck, 2019; Dutriaux & Gyselinck, 2016). However, several other studies have systematically failed to find support for motor affordances in working memory using a variety of experimental paradigms (Canits, Pecher, & Zeelenberg, 2018; Pecher, 2013; Pecher et al., 2013; Quak, Pecher, & Zeelenberg, 2014), leading to a mixed picture.

In a critical review of studies on the role of motor simulations in working memory, Zeelenberg and Pecher (2016) note that many of the paradigms that have yielded results consistent with a functional role of motor simulations in working memory do not in fact provide strong evidence for this claim, because the paradigm itself emphasized actions (e.g., by showing grasping movements before the to-be-remembered objects). They conclude that replications of those studies that provide the most convincing evidence are necessary. Indeed, the value of conducting so-called *direct* replications "intended to evaluate the ability of a particular method to produce the same results upon repetition" (Zwaan, Etz, Lucas, & Donnellan, 2018, p. 5) has recently been emphasized as an important way to make scientific progress by establishing which findings are robust (Munafò et al., 2017; Open Science Collaboration, 2015; Zwaan et al., 2018). Such direct replications are even more important in fields like embodiment that attract intense theoretical debates, because rates of false positives are necessarily increased in such fields (Ioannidis, 2005). We therefore chose to conduct a direct replication of one of the studies that "provide the strongest evidence to date for the view that motor simulations support short-term memory" (Zeelenberg & Pecher, 2016, p. 183).

In a study published in *Cortex*, Shebani and Pulvermüller (2013, SP13 hereafter) presented a striking demonstration of the functional role of the motor system for keeping action words in working memory. Participants had to memorize groups of four words that denoted either arm-related actions (e.g., *peel*, *bash*, *chop*, *clap*) or leg-related actions (e.g., *stomp*, *leap*, *jog*, *hop*). During a six-second memorization phase, they had to carry out a demanding rhythmic pattern (a "paradiddle" drumming drill) at their speed limit with either their arms or legs. Then they had to repeat the four words in the same order they were presented (Figure 1). The results showed a cross-over interaction effect indicating that arm and leg movements led to category-specific memory interference: Arm movements led to more errors recalling arm- than leg-related words, while leg movements led to more errors recalling leg- than arm-related words (Figure 1 inset).

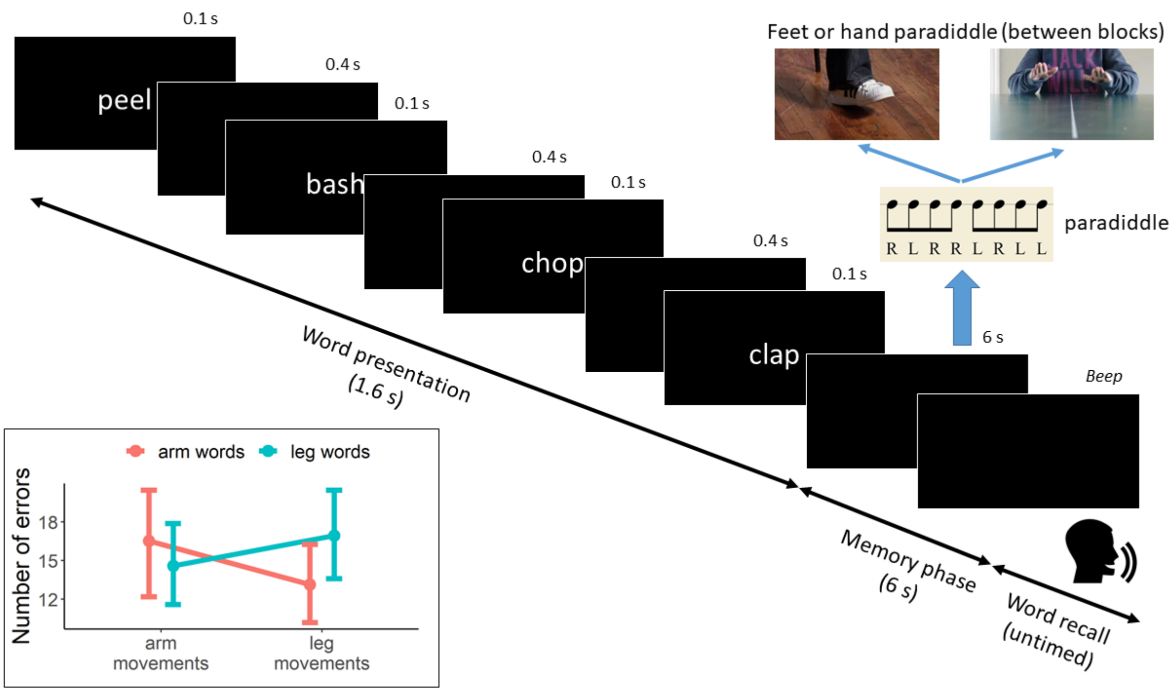

Figure 1. Trial structure and experimental design in SP13; inset figure shows original results. In each trial, participants saw a sequence of four different words that were either all arm-related or leg-related (between trials, within blocks). Words were shown for 100 ms with a stimulus onset asynchrony of 500 ms. Immediately after the offset of the fourth word, participants had to perform a paradiddle (a drumming exercise in which the right [R] and left [L] hands/feet are tapped alternatively and regularly following the pattern RLRRLRLL…) for 6 seconds, either with their hands or with their feet (between blocks, within subjects). After 6 seconds, a beep prompted participants to stop performing the paradiddle and orally repeat the four words in the same order they had seen them. Each block consisted of 24 trials: 12 arm-related and 12 leg-related trials. Inset figure shows the cross-over interaction in the original study based on the data shared by the authors (error bars show non-parametric 95% confidence intervals).

What makes SP13's findings particularly compelling is that they are analogous to a double dissociation in neuropsychology. This allows for a strong inference scheme that attributes a *causal* role to the motor system in working memory, because engaging the part of the motor cortex necessary for arm movements during the arm paradiddle selectively impaired memory for arm-related words, and mutatis mutandis for foot movements and foot-related words. In addition, the fully within-subjects and within-items design (all participants carried out the memory task with the same set of action words twice, once under hand and once under foot interference) means that participants and items served as their own controls. The elegant design and clear-cut results led the authors to conclude that their study was “the first to demonstrate processing impairments critically depending on the meaning of action words as a result of motor system engagement” (Shebani & Pulvermüller, 2013, p. 227).

While the finding in SP13 is of high theoretical relevance, there are also shortcomings that limit the conclusions we may draw from it. A first issue is that the direction of the effect found in SP13 (i.e., that verb-effector congruency would lead to memory *interference* rather than facilitation) was not theoretically predicted beforehand. The authors acknowledge that they “do not fully understand what influences the sign of the effect (facilitation or interference) of motor-language interaction” (SP13, p. 228). Making directional predictions has recently been identified as one of the key challenges for embodiment research (Ostarek & Huettig, 2019). In the absence of such predictions, one pattern of results and its converse might both be taken as support for the same hypothesis, reflecting weak predictive power of the theory.

Further undermining the strength of the initial evidence, a similar later study by the same authors found equivocal results (Shebani & Pulvermüller, 2018). In that study, participants also memorized series of arm- and leg-related words, but this time they had to simply tap their index fingers or their feet while memorizing, instead of carrying out a complex rhythmic pattern as in SP13. In this setting the results showed that participants made *fewer* errors on hand words than leg words in the arm movement (finger-tapping) condition – *prima facie* a facilitation effect. Together with the results in SP13, the authors conjectured that simple, semantically congruent body movements like tapping one’s finger lead to facilitation, whereas complex movements like the hands paradiddle lead to interference (Shebani & Pulvermüller, 2018). However, this interpretation is undercut by the fact that no facilitation effect was found in the foot tapping condition. Instead, the same numerical tendency (fewer errors on hand than leg words) was found when participants tapped their feet, which if anything suggests interference. Crucially, there was no interaction between effector (hand or foot tapping) and verb semantics (arm- or leg-related verbs).[1] The lack of an interaction effect in any direction in a very similar paradigm casts some doubt on the robustness of the initial result.

Another motivation for replicating SP13 is that their conclusions are based on a sample size of only 15 participants, which likely resulted in low statistical power to detect an effect. Increased statistical power is a crucial ingredient for improving replicability in psychological science (Cohen, 1988; Open Science Collaboration, 2015; Zwaan et al., 2018). Unfortunately, low power not only decreases the sensitivity to find a true effect (Cohen, 1988): it also reduces the certainty that a nominally significant finding actually reflects a true effect (Button et al., 2013; Ioannidis, 2005) and leads to exaggerated estimates of those effects (Vasishth, Mertzen, Jäger, & Gelman, 2018). SP13 report an effect size of Cohen's d = 1.25 (p. 226), which is more than 1.5 times larger than what is standardly considered a "large" effect, namely d = 0.8 (Cohen, 1988). However, as detailed below we were not able to reproduce this effect size when re-analysing the original data (Appendix B). Our re-analysis instead suggests that the effect size was very small. In their influential article, Simmons and colleagues recommended to reviewers that "Underpowered studies with perfect results are the ones that should invite extra scrutiny" (Simmons, Nelson, & Simonsohn, 2011, p. 1363). SP13 might be an example of such a study, thus warranting an appropriately powered replication.

[1] The authors report an interaction effect between the hand movement and the control (no movement) conditions (Shebani & Pulvermüller, 2018, p. 5). This is a peculiar choice, given that this comparison was not reported in the original study. Importantly, it does not provide evidence for the double dissociation that makes SP13 so compelling.

Finally, SP13 analysed their error count data using ANOVAs and t-tests, which has several drawbacks that may lead to unreliable statistical inference about the effects of interest (Jaeger, 2008). First, ANOVAs and t-tests assume that the data is continuous and unbounded, but the number of errors in SP13's task is a discrete quantity with upper and lower bounds: For any given four-word trial, the number of errors is bound between 0 and 4; for a block, the upper bound becomes four times the number of trials. The probability model underlying ANOVAs and t-tests can thus erroneously assign probability mass to impossible values beyond the bounds. Furthermore, the variability in error count data depends on the underlying probability of an error: It is largest for probabilities close to 0.5 and smaller for probabilities close to 0 and 1 (Jaeger, 2008). This violates the homoscedasticity assumption of ANOVAs and t-tests. A better choice – and also the one we adopt here – is to analyse the data with generalized linear mixed models, as the probability model underlying this analysis is well suited for error count data (Jaeger, 2008). Additionally, subject- and item-level variability can simultaneously be modelled, leading to improved inferences about population-level effects (Baayen, Davidson, & Bates, 2008; Gelman & Hill, 2007).

Our aim is thus to run a direct replication of SP13, pre-registering all aspects of data collection and data analysis, and introducing only minimal changes to the original design (detailed below). We seek to replicate the finding that executing arm or leg movements selectively impairs working memory for arm- and leg-related action words, respectively. This constitutes a strong test of the claim that the sensorimotor system shares processing resources with working memory for action words and thus "can be considered to be *necessary* for action-word memory" (SP13, p. 227, emphasis in original). To plan for compelling evidence, we adopt a prospective sequential Bayes factor design analysis (Schönbrodt & Wagenmakers, 2018). In our replication, we set the minimum sample size to N=60 (four times that of the original) and the maximum to N=96 (over six times the original), with step sizes of 12 participants.[2] We clearly define a stopping rule for data collection based on a pre-determined threshold as to what constitutes evidence for or against the alternative hypothesis using Bayes factors (Dienes, 2014; Verhagen & Wagenmakers, 2014). Our design ensures statistical power above 90% even if the effect size was 2.5 times smaller than reported in SP13 (see Sample size rationale below). The analysis improves on the original by modelling the data with a Bayesian generalized linear mixed model suitable for error data and by using Bayes factors to quantify the results of the replication attempt (Verhagen & Wagenmakers, 2014).

2. Method

Figure 1 shows the design used in SP13; we refer the reader to the original study for additional details. We contacted the authors regarding aspects of the design that remained unclear from their report and will follow their clarifications unless otherwise stated. Below we report the methods making explicit any divergence from the original. Appendix A provides a systematic comparison of our replication and the original, following Brandt et al.'s (2014) "replication recipe".

2.1 Participants and sample size rationale

2.1.1 Original study

SP13 recruited data from 15 monolingual native speakers of English (8 males) aged 18–30. Participants were right-handed, reported normal vision and hearing, and had no history of neurological or psychiatric illness. Musicians were excluded from the experiment.

[2] Our maximal sample size is almost 3.5 times larger than the median sample size of the 31 experiments in the 11 studies we reviewed on working memory and motor interference (median: 28; range: 16–52).

2.1.2 Our replication

Our study will be conducted in Sweden and we will therefore recruit native speakers of Swedish in the same age range as the original (18–30). We will adopt a sequential Bayes factor design (Schönbrodt & Wagenmakers, 2018) with a minimum sample size of 60 and a maximum sample of size 96 participants with step sizes of 12 participants.[3] Participants excluded from the statistical analysis due to pre-specified exclusion criteria (see below) will be replaced by new participants and the number of exclusions will be reported. The exact sampling plan is shown in Figure 2 and its rationale is given in the following section.

1. Collect data from minN=60 participants (four times the sample size of the original).
2. Compute the BF with a weakly informative prior (see Analyses below).
3. If $BF_{10} > 6$ or $BF_{01} > 6$, stop data collection and report results. Else:
4. If $N < 96$, collect another batch of 12 participants and go to step 2. Else:
5. When we reach maxN=96, stop data collection, compute BFs and report results.

Figure 2. Sequential design of the replication.

As in the original, we will screen participants for right-handedness, normal vision and hearing, and lack of history of neurological or psychiatric illnesses. We will exclude musicians, operationalized as anybody who has at least five years of formal musical training or equivalent informal experience. We will also exclude participants who report having played the drums for more than one year. Monolingual Swedish speakers are virtually impossible to find in the

[3] Choosing N that is a multiple of 12 enables counterbalancing of the three stimuli lists, the order in which they are encountered, and the order of the interference blocks (hands first or feet first).

targeted age range and educational level, as English language instruction is compulsory in Swedish education and communicative English proficiency is generally high (Bylund & Athanasopoulos, 2015; Skolverket [Swedish National Agency for Education], 2011). We therefore adopted the following standard definition for who counts as a native speaker and may therefore participate in the study (cf. Abrahamsson & Hyltenstam, 2009; Bylund, Abrahamsson, Hyltenstam, & Norrman, 2019): Participants should a) be born in Sweden, b) be exposed to Swedish since birth and without significant interruption (i.e., not more than six months) throughout their lives; c) have grown up in a Swedish-speaking home; and d) have Swedish as their dominant language.

2.1.3 Sample size rationale

We adopted a prospective Bayes factor design analysis to plan sample size (Schönbrodt & Wagenmakers, 2018). In contrast to *p* value-based inference, using Bayes factors allows for a 3-way decision once the data are collected, based on pre-specified evidence thresholds: The data may a) support the alternative hypothesis that there is an effect, b) support the null hypothesis that no effect exists, or c) remain inconclusive (Dienes, 2014; Wagenmakers, 2007). The goal then is to design a study that jointly yields a high probability of obtaining strong evidence (i.e., data that do not remain inconclusive) and minimizes the probability of misleading evidence (i.e., data that lead to accepting the wrong hypothesis) (Schönbrodt & Wagenmakers, 2018). This framework makes it possible to implement a sequential design that pre-specifies a minimum sample size (minN), a plan to test additional batches of participants if the required degree of evidence is not reached at a given sample size, and a maximum sample size (maxN) at which for practical considerations data collection stops, irrespective of the degree of evidence reached.

To determine minN and maxN for our study, we used the state-of-the-art R package *BFDA* (Schönbrodt & Stefan, 2018), which allows researchers to simulate studies under different statistical assumptions and summarize the results. Following *Cortex* guidelines, in all simulations we set the threshold for accepting the alternative over the null hypothesis (or vice versa) at a Bayes factor of 6 ($BF_{10} > 6$ or $BF_{01} > 6$). We used non-informative default priors so as to make our estimates more conservative (Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019). We next synthesize the outcome of our simulations and how they justify the current design (see Appendix C for details).

The two critical parameters for the simulations are the estimated effect size and the type of analysis. The original study reported an effect size of Cohen's $d = 1.25$ for the critical interaction in a 2-by-2 within-subjects ANOVA (SP13, p. 226). This effect size is substantially larger than a "large" effect ($d = 0.8$) according to Cohen's (1988) practical guidelines. Using this effect size estimate with a paired t-test design in *BFDA*, our minimal sample size minN=60 would guarantee detecting the effects in 100% of simulated studies, with a 0% of false-negative or inconclusive results. In fact, with such an effect size a sample size of just N=12 would have 94% power (Appendix C, sect. 4).

There are, however, potential problems with this effect size estimate. We first reanalyzed the original data shared by the authors with the same analysis type, repeated measures ANOVAs (Appendix B, sect. 4).[4] Although we could precisely reproduce their summary statistics and the *F* and *p* values for their ANOVAs, suggesting that there are no issues with the data set itself, we were unable to reproduce their effect size estimate of Cohen's $d = 1.25$. Second, the suboptimal analysis method (ANOVA) used in SP13 might have led to an unreliable effect size estimate. A

[4] The original data are available at https://zenodo.org/record/3402035#.XZjAJkb7RaQ.

more appropriate analysis would treat the error counts as arising from a binomial distribution. When we re-analyzed the original data with this improved analysis method (Appendix B, sect. 6), the critical interaction effect was still robustly there and different from zero (Bayes factor = 12.1 against the null hypothesis). However, the effect size estimate became very small, namely 0.15 log-odds. Note that the unit for effect size in binomial models (log-odds) is not on the same scale as Cohen's *d*: the estimated effect (0.15 log-odds) is less than a third of what is considered a "small" effect, namely 0.5 log-odds (Chen, Cohen, & Chen, 2010).

While the BFDA does not allow for fine-tuning the exact intended experimental design and analyses, simulations can be run with the more appropriate "AB-test" method, which implements Bayes factors for testing the equality of binomial proportions (Kass & Vaidyanathan, 1992; Schönbrodt & Stefan, 2018). If we use the posterior distribution of the critical interaction effect from our improved reanalysis as the input to the effect size estimate in our simulations, then not even a sample size as large as N=3000 participants would ensure enough power to detect the effect (Appendix C, sect. 5). Even with such an enormous sample size, only 48% of studies would lead to correctly accepting the alternative hypothesis, while 46% would lead to incorrectly accepting the null (only 5% would yield inconclusive evidence). In sum, based on these simulations, planning for compelling evidence with an effect size that small is practically unfeasible as it would require sample sizes in the high thousands.

To nevertheless test the claim put forward by SP13 that arm and leg movements selectively interfere with working memory for arm and leg words, respectively, we opted for a sequential design with a maximum N=96 that would send an important message to the field even if the results remained inconclusive, in line with this journal's guidelines. Our maximal sample size is more than six times that in the original and more than three times that of typical studies testing effects of motor affordances in working memory (see footnote 2). More importantly, our design ensures very high power of 98% to detect a “medium” effect size of 1.25 log-odds (equivalent to Cohen’s $d = 0.5$, Chen et al., 2010), which is 2.5 times smaller than the effect size reported in SP13; it remains high-powered (83%) to detect a “small”-to-“medium” effect size of 0.9 log-odds (equivalent to Cohen’s $d = 0.36$), which is 3.5 times smaller than that reported in SP13. Figure 3 illustrates estimated power under a “medium” effect size of 1.25 log-odds: 98% of the studies lead to a correct detection of the effect of interest, 0% lead to false-negatives, and 2% remain inconclusive (Appendix C, sect. 6.2). Under a “small”-to-“medium” effect size (0.9 log-odds), the effect would be correctly detected in 83% of the cases, there would be 0% false negatives, and 17% of inconclusive results (Appendix C, sect. 6.3). Note that we deliberately used uninformative priors in all simulations, which therefore represent conservative power estimates compared to priors that would take into account the previous positive results (Verhagen & Wagenmakers, 2014).

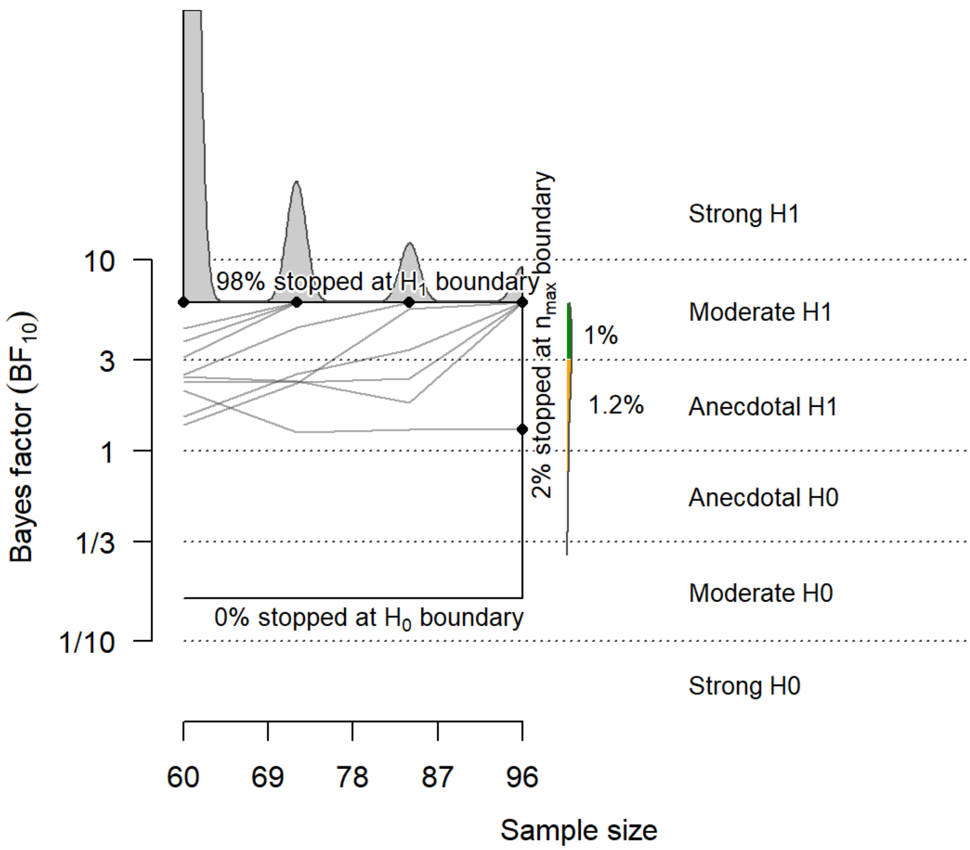

Figure 3. Sequential Bayes factor (BF) design with maximum N=96 assuming a "medium" effect size (1.25 log-odds) for the critical interaction. The figure shows a summary of 10000 simulated studies starting with a minN=60 and adding batches of 12 participants if the results do not reach the intended evidence threshold (BF>6), until reaching maxN=96. We used a default (non-informative) analysis prior and the "AB-test" as analysis method. Simulations were run with the *BFDA* package.

2.2 Materials

2.2.1 Original study

SP13 used 36 arm-related and 36 leg-related English verbs as their stimuli. The words in the two lists were matched for a range of psycholinguistically relevant variables. Critically, the two lists differed significantly on arm-relatedness (arm words: 5.46 [SE=0.14]; leg words = 1.92 [0.12]) and leg-relatedness (arm words: 2.28 [0.13]; leg words = 5.58 [0.22]), as assessed by semantic ratings (the scale is not reported in SP13).

2.2.2 Our replication

To increase statistical power (Brysbaert & Stevens, 2018), we will increase the number of items to 52 arm-related and 52 leg-related Swedish verbs, which is the largest set of words we could find while keeping the two lists of equal length and matching them along the same psycholinguistic variables as in the original: Number of letters, number of phonemes, word frequency, grammatical ambiguity, lemma frequency, bigram frequency, trigram frequency, valence, arousal, and imageability (see Table 1).[5] Crucially, our two lists also differed significantly on arm-relatedness (arm words: 6.59 [SE=0.03]; leg words = 1.80 [0.07]) and leg-relatedness (arm words: 1.34 [0.03]; leg words = 6.46 [0.08]), as assessed by semantic ratings on a 7-point scale obtained from 12 Swedish native speakers. See Appendix D1 for the full list of stimuli and Appendix D2 for an explanation of how each variable was computed.

Table 1. Means, standard errors and *p* values (from unpaired t-tests) comparing psycholinguistic variables of the 52 arm and 52 leg words used in this study.

	Arm words		Leg words		
Feature	Mean	SE	Mean	SE	*p* value (t-test)
Number of letters	5.13	0.13	5.37	0.18	0.3
Number of phonemes	4.69	0.1	5.02	0.16	0.1
Word log frequency	2.56	0.09	2.28	0.13	0.1
Lemma log frequency	2.79	0.09	2.62	0.13	0.3
Bigram log frequency	6.02	0.04	6.03	0.05	0.8
Trigram log frequency	4.82	0.07	4.84	0.07	0.8
Grammatical ambiguity	0.2	0.02	0.16	0.02	0.2
Valence	3.67	0.1	3.79	0.11	0.4
Arousal	2.49	0.09	2.32	0.09	0.2
Imageability	5.54	0.06	5.33	0.1	0.1
Arm-relatedness	6.59	0.03	1.8	0.07	<.001
Leg-relatedness	1.34	0.03	6.46	0.08	<.001

[5] Since the original study did not explain how some of these measures were obtained, we contacted the authors and operationalized the variables based on this correspondence. We omitted three of the original variables (visual relatedness, body relatedness, and general action relatedness) that were redundant with other collected measures according to the authors (F. Pulvermüller, personal communication, May 30, 2019).

2.3 Procedure

2.3.1 Original study

The basic procedure is shown in Figure 1. Each trial began with a fixation point shown in the centre of the screen for 3 seconds. After this, the four words of the trial (all either arm- or leg-related) were presented serially. Each word was presented for 100 ms with a 500 ms stimulus onset asynchrony. Presentation of the fourth word was followed by a 6 second memory phase during which participants had to retain the four words in memory in the same order as they were presented. The memory phase ended with a beep which prompted participants to repeat the four words in the order they had encountered them. SP13 used two pseudo-randomized stimulus sequences, counterbalanced across subjects. The order of arm-word trials and leg-word trials within a block was randomized with the constraint that not more than three trials of the same word type appeared consecutively.

In the two critical conditions (hand and foot movement), participants had to carry out a drumming exercise known as the “paradiddle”, in which the right (R) and left (L) hands/feet are tapped alternatively and regularly following the pattern RLRRLRLL, etc. The motor task was made more challenging by having participants carry out the memory task while performing the paradiddle at their frequency threshold. This threshold was determined for each individual participant before the beginning of the relevant block (hand or foot interference) of the memory task, as follows: After getting familiarized with the basic form of the paradiddle, participants started performing it at 100 beats per minute using a metronome. The experimenter gradually increased the frequency by 10 beats if participants were able to perform the paradiddle without errors for 20 seconds. Each participant’s hand/foot frequency threshold was defined as the highest pace at which they could maintain error-free performance for 20 seconds. In addition to the two critical interference blocks (hand and foot movement), SP13 had a control condition, in which participants were asked to keep silent during the 6 second memory phase, and an articulatory condition, in which participants had to repeat the syllable *bla* throughout the memory phase. These will not be further discussed as we will not include them in our replication (see below and Appendix A).

Trial presentation was self-paced and initiated by pressing the space bar. Written and oral instructions were given before each block. Participants were offered ample opportunity to practice before starting a block and were allowed to take breaks between blocks and between trials.

One aspect that remains ambiguous from the original report is the exact number of trials per block. SP13 first indicate that there were “twenty-four trials in each block, twelve arm-word trials and twelve leg-word trials” (p.225). However, later in the same paragraph they note that “the full set of 72 words [was] presented twice in all conditions”. Both cannot be right since presenting 72 words twice (i.e., 144 words à four words per trial) would amount to 36 trials. We checked with the authors who clarified that the former figure (24 trials) was the correct one, noting that “48 words from each category were shown in each block. Twelve words per category, randomly selected, were repeated once in each block” (Z. Shebani, personal communication, April 1, 2018).

2.3.2 Our replication

Our replication exactly follows the procedure reported in SP13 with the following exceptions. First, we will only include the two critical conditions – the hand-movement and foot-movement conditions – and omit the control and articulatory conditions. The latter two are orthogonal to the hypothesis being tested and treated as such in SP13, who consistently refer to the hand- and foot-movement conditions as the “critical conditions […] directly addressing the main hypothesis motivating this study" (p. 225–226). Importantly, there is no theoretical reason to assume that the embodiment effect depends on participants also performing a control and articulatory suppression conditions. And since the "order of the blocks was counterbalanced across subjects using a Latin-square design" (SP13, p. 225), the critical effects were made orthogonal to the non-critical conditions by design.

Secondly, we assume that repeating a random subset of 12 out of 48 words per block (as in the original) was not critical to the obtained result and thus opt for a more standard design in which each word is shown once per block. Since we have a larger set of stimuli, this still results in more trials per block than in SP13 (26 instead of 24).

Third, we will use three (rather than two) random lists grouping the stimuli words into different 4-word items.[6] Each participant will see two different groupings/lists (one per block), with lists, order of blocks (hand–leg movements / leg–hand movements) counterbalanced across participants (Appendix F). The specific order in which the items of a list are shown is random for each participant-block while respecting the original constraint that there appear no more than three consecutive trials of the same word type.

Finally, we will perform an additional check during data collection: We will systematically monitor if participants make errors on the paradiddles during the memory task, so that we can use this information to exclude participants who systematically fail to carry out the rhythmic task (see Exclusion criteria below).[7]

[6] The original authors clarified that they used two pseudo-randomized lists (Z. Shebani, personal communication, April 1, 2018), but the exact lists could not be made available.
[7] The original authors clarified that "Mistakes in paradiddles were not monitored/recorded as accuracy in performing the paradiddles was not the focus of the study" (Z. Shebani, personal communication, April 1, 2018). While we agree that the number of rhythmic errors is not the focus of interest, we believe that carrying out the paradiddle is a prerequisite to test the critical hypothesis.

2.4 Data exclusion criteria

The experimenter will monitor the execution of the rhythmic task (hands/feet paradiddle) throughout the session using an error monitor form. The following will be considered execution errors: forgetting to execute the paradiddle during the memory phase or executing the paradiddle with the wrong effector. Trials with execution errors will be excluded from the analysis. At the participant level, we will exclude participants who fail to carry out the interference task in more than 30% of the trials across blocks or in more than 50% of trials in a single block. We will also exclude participants for which, due to technical failure (e.g., recording not working), data is missing for more than 30% of the trials across blocks or for more than 50% of trials in a single block. All exclusions will take place before the data is coded and analysed. Excluded participants will be replaced.

2.5 Quality checks

As a quality check we will verify that there are no ceiling or floor effects (i.e., 0% or 100% errors) in any of the experimental cells defined by the 2 (word type) x 2 (movement interference) design. Ceiling/floor effects are not expected given the original results and our own piloting of the basic memory task (errors on 15-40% of trials).

As a positive control we will analyse the effect on recall of serial position of a word within a trial. Serial position effects are among the most robust effects in working memory research (see Popov & Reder, 2019 for a recent review). This check is orthogonal to our main hypothesis and merely serves as an outcome-neutral criterion to verify that we can replicate a pervasive effect in working memory tasks and that participants were engaged. This effect was present in our own pilot of the basic task (without interference) with 17 participants (estimate = 0.39 log-odds, SE = 0.032, $p < .001$; analysed using logistic mixed model regression). Thus both expert judgement (V. Popov, personal communication, September 23, 2019) and our own pilot suggest this effect is virtually guaranteed to appear in the data. We will test this effect by fitting a logistic mixed model to the data with recall error as the binary dependent variable (0=word remembered, 1=word not remembered) and the following fixed-effect predictors (all predictors standardized): word position within trial, trial position within the experiment, error on any of the preceding words in trial (binary), word type, movement interference. Random effects will include a by-participant random intercept and random slope for word position within trial, as well as a random intercept by verb.[8] For this analysis, we will use the R package *lme4* (Bates, Mächler, Bolker, & Walker, 2015).

3. Data coding and analyses

3.1 Data coding

We will adopt a binary coding for the data: For each word within a 4-word memory trial, the dependent variable will be 1 if the verb is recalled and 0 if it is not. Thus, there will be four observations per trial and 104 observations per participant-block (52 of each word type).

Our coding differs from that of SP13 in that it disregards shift errors, an error type whose removal did not affect the critical interaction effect and that accounted for 12% of all errors in the original (SP13, p. 226). To understand shift errors, consider a trial that consists of the words *peel-bash-chop-clap*. If the participant response is *bash-peel-chop-clap*, this will be counted as zero errors according to our coding but it would be counted as one error (a shift error) in SP13's coding scheme, because the order of *peel* and *bash* is interchanged. We opted for this divergence for several reasons. First, we did not obtain an algorithm from the authors that would allow us to unambiguously reproduce their error coding scheme from a written transcription of participant

[8] Model formular in R: Error ~ word_in_trial_z + trial_in_experiment_z + preceding_error_in_trial + word_type + movement_interference_condition + (1 + word_in_trial_z | subj) + (1 | verb)

responses. SP13 report three types of errors: omissions, replacements, and shifts (they also mention that additions counted as errors [p. 225] but do not report the rate of this error type). Some coding decisions are inherently arbitrary; for example, a replacement (one error) could equally well be coded as an omission and an addition (two errors). For want of a principled protocol that can be implemented in a machine, we preferred to adopt our more transparent coding scheme. Second, counting shift errors just as any other error type makes the underlying assumption that all error types carry the same weight, which can lead to counterintuitive outcomes. For example, a participant response such as *bash-clap-peel-chop* for the trial above (where all words are correctly remembered) would count as three errors (three shifts), exactly the same as if the response had been *peel-potato-garden-I don't know* (two replacement errors and an omission). Intuitively, it would seem that the former response reflects superior memory than the latter, but this would not be captured by the coding. Third, as already mentioned, none of the critical results reported in SP13 hinged on shift errors: SP13 report that the critical interaction was still present if shift errors were removed and that it was not present if these errors were evaluated separately (SP13, p. 226). Finally, our coding scheme allows for improved inference on population-level effect estimates by letting us model item variability as a random effect. This is straightforward when each binary response can be linked to a specific verb (as in our coding), but it becomes difficult in the case of shift errors.[9]

3.2 Inter-rater reliability

Initially 5% randomly selected observations from the first 60 participants (i.e., 624 data points) will be transcribed and coded independently by two raters who are native speakers of Swedish. If

[9] We note that it is easy from our transcripts to implement an alternative coding scheme in which all error types (including shifts) are counted (e.g., by computing Levensthein distance from the string provided by the participant to the target string, where each word counts as a symbol). However, for the above reasons such a coding will not be the basis for our primary pre-registered analysis.

the inter-rater agreement is ≥ 95%, each of the raters will proceed to code separate subsets of the complete data set. If inter-rater agreement is < 95%, disagreements will be inspected and resolved through discussion, so that coding criteria become shared among raters. Then a separate 5% sample of the data will be coded and the procedure repeated until inter-rater agreement is ≥ 95%. The number of rounds needed and the inter-rater agreement at each round will be reported.

3.3 Analytic approach: Bayesian logistic mixed effects regression

We will analyse the data using a Bayesian version of logistic mixed effects regression implemented in the package *brms* (Bürkner, 2017) in the R statistical environment (R Core Team, 2015). Logistic mixed effects regression is well suited to model binary outcomes and relies on the log of the odds as a link function (see Jaeger, 2008). The dependent binary variable Error (=1 if a word is missed, =0 if it is remembered; see Data coding) will be modelled as a function of the contrast-coded predictors Interference Movement (1=arm movements, −1=leg movements), Word Type (1=arm-related words, −1=leg-related words), and their interaction. To determine the random effect structure of the model, we will follow the guidelines in Barr, Levy, Scheepers, and Tily (2013): We will start by fitting the maximal model justified by the design, which here corresponds to by-participant random intercepts and random slopes for Movement, Word Type, and their interaction, as well as by-item random intercepts and random slopes by Movement. In case of sampling problems during the model fitting procedure, we will simplify this random effect structure in the principled way outlined in Appendix G. Additionally we will include the following nuisance variables as fixed effect predictors in the model (centred and scaled): trial position within the experiment, error on any of the preceding words in trial (binary), word position within trial. A full analysis pipeline based on simulated data is available in Appendix E.

In the Bayesian framework, priors need to be specified for all model parameters. We will standardize predictors and then set a weakly informative prior for all coefficients: a Normal distribution centred on zero, with a standard deviation of 2. This corresponds with the prior belief that any given coefficient is likely to be small, while allowing for a coefficient to be larger if the data support it; it is broadly equivalent to (weakly regularizing) ridge regression in the frequentist framework (Mallick & Yi, 2013). For all standard deviations of group-level random effects, we will use the corresponding default priors, which are "used (a) to be only very weakly informative in order to influence results as few as possible, while (b) providing at least some regularization to considerably improve convergence and sampling efficiency" (https://rdrr.io/cran/brms/man/get_prior.html; Bürkner, 2017). See Appendix E for details.

We will report mean estimates and modes, standard errors, and 95% credible intervals for all fixed effects model parameters. The data set and analysis script will be openly shared.

3.4 Stopping rule and assessing the outcome of the replication with Bayes factors

To decide when to stop data collection (see Figure 2 and Participants and sample size rationale) and to make a decision as to whether our replication successfully detects the effect reported in SP13 or fails to do so, we will use Bayes factors (see Dienes, 2014; Verhagen & Wagenmakers, 2014, and references therein). Bayes factors quantify the odds that one among two (or more) hypotheses is true rather than the other(s), given the data. The contrast typically involves an alternative and a null hypothesis. We will compute the following two Bayes factors (see Verhagen & Wagenmakers, 2014):

1. BF1: Independent Jeffreys–Zellner–Siow (JZS) Bayes Factor to address the question *if the effect is present or absent* in the replication attempt.
2. BF2: Replication Bayes factor to address the question *if the "effect from the replication attempt [is] comparable to what was found before, or [is] absent*?" (Verhagen & Wagenmakers, 2014, p. 1458).

What differs between BF1 and BF2 is how much weight is given to the previous results obtained in SP13: BF1 does not take them into account (weakly informative prior on interaction effect: N(0, $\sigma = 2$)), while BF2 uses as prior a normal distribution based on the posterior estimates of the model fitted to the original data.

Our decision as to when to stop data collection (see Figure 2) will be based on the calculation of BF1 only. Once data collection has stopped (either because BF1>6 in favour of one of the competing hypotheses or because we have reached maxN=96) BF2 will be computed.

Both BFs will be reported. A clear replication success will be an outcome in which both $BF1_{10} > 6$ and $BF2_{10} > 6$. Conversely, a clear failure to replicate will be an outcome in which $BF1_{01} > 6$ and $BF2_{01} > 6$. If only one of the two BFs reach the targeted threshold, our primary interpretation will be based on BF1, but it will be nuanced by the outcome of BF2. The value of BFs will be interpreted according to the heuristics in Table 2.

Table 2. Heuristic classification scheme for the interpretation of Bayes factors BF_{10} (adjusted from Schönbrodt & Wagenmakers, 2018). The same scheme will be used to interpret BF_{01}.

Bayes factor	Evidence category
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
6 – 10	Evidence for H_1
3 – 6	Anecdotal evidence for H_1
1 – 3	Inconclusive evidence

Declarations of interest

None.

List of appendices

- Appendix A: Systematic comparison of the original study and our replication following Brandt et al.'s (2014) "replication recipe".
- Appendix B: Reanalysis of the original data.
- Appendix C: Bayes factor design analysis
- Appendix D1: List of stimuli with measures on lexical and psycholinguistic variables
- Appendix D2: Explanation of variables in Appendix D1
- Appendix E: Analysis pipeline
- Appendix F: Counterbalancing of lists across participants
- Appendix G: Algorithm for model simplification in case of sampling issues during model fitting

Note: All appendices and the necessary code to reproduce all analyses in Appendices B, C, and E can be found at https://osf.io/ktsfw/?view_only=63e3071ba35641a0ba11785324e427e3

Acknowledgements

We would like to thank T. Florian Jaeger (and the HLP lab) and Mante Nieuwland for helpful suggestions, and Pia Järnefelt and Margareta Majchrowska for help with practical preparations. This work was supported by the Swedish Research Council [grant numbers 2015-01317 to Emanuel Bylund and 2018-00245 to Guillermo Montero-Melis].

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126. https://doi.org/10.1016/S1364-6613(00)01593-X

Aziz-Zadeh, L., & Damasio, A. (2008). Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology-Paris*, *102*(1), 35–39. https://doi.org/10.1016/j.jphysparis.2008.03.012

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

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., & Dale, H. C. A. (1966). The effect of semantic similarity on retroactive interference in long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *5*(5), 417–420. https://doi.org/10.1016/S0022-5371(66)80054-3

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). https://doi.org/10.1016/S0079-7421(08)60452-1

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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. *Journal of Cognitive Neuroscience*, *18*(10), 1607–1615. https://doi.org/10.1162/jocn.2006.18.10.1607

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van ’t Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

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

Bylund, E., Abrahamsson, N., Hyltenstam, K., & Norrman, G. (2019). Revisiting the bilingual lexical deficit: The impact of age of acquisition. *Cognition*, *182*, 45–49. https://doi.org/10.1016/j.cognition.2018.08.020

Bylund, E., & Athanasopoulos, P. (2015). Televised Whorf: Cognitive Restructuring in Advanced Foreign Language Learners as a Function of Audiovisual Media Exposure. *The Modern Language Journal*, *99*(S1), 123–137. https://doi.org/10.1111/j.1540-4781.2015.12182.x

Canits, I., Pecher, D., & Zeelenberg, R. (2018). Effects of grasp compatibility on long-term memory for objects. *Acta Psychologica*, *182*, 65–74. https://doi.org/10.1016/j.actpsy.2017.11.009

Chen, H., Cohen, P., & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864. https://doi.org/10.1080/03610911003650383

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Downing-Doucet, F., & Guérard, K. (2014). A motor similarity effect in object memory. *Psychonomic Bulletin & Review*, *21*(4), 1033–1040. https://doi.org/10.3758/s13423-013-0570-5

Dutriaux, L., Dahiez, X., & Gyselinck, V. (2019). How to change your memory of an object with a posture and a verb. *Quarterly Journal of Experimental Psychology*, *72*(5), 1112–1118. https://doi.org/10.1177/1747021818785096

Dutriaux, L., & Gyselinck, V. (2016). Learning Is Better with the Hands Free: The Role of Posture in the Memory of Manipulable Objects. *PLOS ONE*, *11*(7), e0159108. https://doi.org/10.1371/journal.pone.0159108

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: The role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3/4), 455–479. https://doi.org/10.1080/02643290442000310

García, A. M., & Ibáñez, A. (2016). A touch with words: Dynamic synergies between manual actions and language. *Neuroscience & Biobehavioral Reviews*, *68*, 59–95. https://doi.org/10.1016/j.neubiorev.2016.04.022

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*(1), 1–19. https://doi.org/10.1017/S0140525X97000010

Guérard, K., Guerrette, M.-C., & Rowe, V. P. (2015). The role of motor affordances in immediate and long-term retention of objects. *Acta Psychologica*, *162*, 69–75. https://doi.org/10.1016/j.actpsy.2015.10.008

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

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. *Language and Cognitive Processes*, *25*(6), 749–776. https://doi.org/10.1080/01690961003595572

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 129–144. https://doi.org/10.1111/j.2517-6161.1992.tb01868.x

Lagacé, S., & Guérard, K. (2015). When motor congruency modulates immediate memory for objects. *Acta Psychologica*, *157*, 65–73. https://doi.org/10.1016/j.actpsy.2015.02.009

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, *30*(4), 420–429. https://doi.org/10.1080/23273798.2014.987791

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

Mallick, H., & Yi, N. (2013). Bayesian Methods for High Dimensional Linear Models. *Journal of Biometrics & Biostatistics*, *1*, 005. https://doi.org/10.4172/2155-6180.S1-005

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547. https://doi.org/10.1098/rsos.150547

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. https://doi.org/10.1038/s41562-016-0021

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science, 4*(2), 135–183. https://doi.org/10.1207/s15516709cog0402_2

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

Ostarek, M., & Huettig, F. (2019). Six Challenges for Embodiment Research. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721419866441

Pasternak, T., & Zaksas, D. (2003). Stimulus Specificity and Temporal Dynamics of Working Memory for Visual Motion. *Journal of Neurophysiology*, *90*(4), 2757–2762. https://doi.org/10.1152/jn.00422.2003

Pecher, D. (2013). No role for motor affordances in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 2–13. http://dx.doi.org/10.1037/a0028642

Pecher, D., Klerk, R. M. de, Klever, L., Post, S., Reenen, J. G. van, & Vonk, M. (2013). The role of affordances for working memory for objects. *Journal of Cognitive Psychology*, *25*(1), 107–118. https://doi.org/10.1080/20445911.2012.750324

Popov, V., & Reder, L. M. (2019). Frequency effects on memory: A resource-limited theory. *Psychological Review*. http://dx.doi.org.ezp.sub.su.se/10.1037/rev0000161

Postle, B. R., Idzikowski, C., Sala, S. D., Logie, R. H., & Baddeley, A. D. (2006). The selective disruption of spatial working memory by eye movements. *Quarterly Journal of Experimental Psychology*, *59*(1), 100–120. https://doi.org/10.1080/17470210500151410

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

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360. https://doi.org/10.1038/nrn2811

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

Pulvermüller, F., Kherif, F., Hauk, O., Mohr, B., & Nimmo-Smith, I. (2009). Distributed cell assemblies for general lexical and category-specific semantic processing as revealed by fMRI cluster analysis. *Human Brain Mapping*, *30*(12), 3837–3850. https://doi.org/10.1002/hbm.20811

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132. https://doi.org/10.1017/S0140525X00002053

Quak, M., Pecher, D., & Zeelenberg, R. (2014). Effects of motor congruence on visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2063–2070. https://doi.org/10.3758/s13414-014-0654-y

R Core Team. (2015). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Raposo, A., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia*, *47*(2), 388–396. https://doi.org/10.1016/j.neuropsychologia.2008.09.017

Schönbrodt, F. D., & Stefan, A. M. (2018). BFDA: An R package for Bayes factor design analysis (Version 0.5.0). Retrieved from https://github.com/nicebread/BFDA

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Shebani, Z., & Pulvermüller, F. (2013). Moving the hands and feet specifically impairs working memory for arm- and leg-related action words. *Cortex*, *49*(1), 222–231. https://doi.org/10.1016/j.cortex.2011.10.005

Shebani, Z., & Pulvermüller, F. (2018). Flexibility in Language Action Interaction: The Influence of Movement Type. *Frontiers in Human Neuroscience, 12*. https://doi.org/10.3389/fnhum.2018.00252

Shtyrov, Y., Butorina, A., Nikolaeva, A., & Stroganova, T. (2014). Automatic ultrarapid activation and inhibition of cortical motor systems in spoken word comprehension. *Proceedings of the National Academy of Sciences*, *111*(18), E1918–E1923. https://doi.org/10.1073/pnas.1323158111

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Skolverket [Swedish National Agency for Education]. (2011). *Internationella språkstudien [The international language survey]* (No. 375). Retrieved from https://www.skolverket.se/publikationer?id=2832

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., … Perani, D. (2005). Listening to Action-related Sentences Activates Fronto-parietal Motor Circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281. https://doi.org/10.1162/0898929053124965

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

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 830–846. http://dx.doi.org.ezp.sub.su.se/10.1037/0096-1523.24.3.830

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, *8*(6), 769–800. https://doi.org/10.1080/13506280042000144

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. http://dx.doi.org/10.1037/a0036731

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

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual Experience Shapes Object Representations. *Psychological Science*, *24*(6), 909–919. https://doi.org/10.1177/0956797612464658

Zeelenberg, R., & Pecher, D. (2016). The Role of Motor Action in Memory for Objects and Words. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 64, pp. 161–193). https://doi.org/10.1016/bs.plm.2015.09.005

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