CORTEX 150 (2022) 108–125

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Registered Report

No evidence for embodiment: The motor system is not needed to keep action verbs in working memory

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

[a] Max Planck Institute for Psycholinguistics, the Netherlands
[b] Department of Linguistics, Stockholm University, Sweden
[c] Centre for Research on Bilingualism, Stockholm University, Sweden
[d] Department of General Linguistics, Stellenbosch University, South Africa

ARTICLE INFO

Article history:
Protocol received: 9 October 2019
Protocol approved: 29 February 2020
Received 3 November 2021
Reviewed 19 December 2021
Revised 27 February 2022
Accepted 28 February 2022
Action editor Chris Chamber
Published online 11 March 2022

Keywords:
Embodiment
Working memory
Semantics
Action verbs
Replication
Registered report

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). Such a striking double dissociation is acknowledged even by critics to constitute strong evidence for an embodied account of working memory. Here, we report on an attempt at a direct replication of this important finding. We followed a sequential sampling design and stopped data collection at $N = 77$ (more than five times the original sample size), at which point the evidence for the *lack* of the critical selective interference effect was very strong ($BF_{01} = 91$). This finding constitutes strong evidence against a functional contribution of the motor system to keeping action verbs in working memory. Our finding fits into the larger emerging picture in the field of embodied cognition that sensorimotor simulations are neither required nor automatic in high-level cognitive processes, but that they may play a role depending on the task. Importantly, we invite researchers to engage in transparent, high-powered, and fully pre-registered experiments like the present one to ensure the field advances on a solid basis.

* *Corresponding author*. Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525, XD Nijmegen, the Netherlands.
E-mail addresses: Guillermo.MonteroMelis@mpi.nl, montero.gui@gmail.com (G. Montero-Melis).

https://doi.org/10.1016/j.cortex.2022.02.006
0010-9452/

1. Introduction

What is the nature of the system underlying high-level cognitive functions in the human brain?[1] 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 verb 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 et al., 2004; Pulvermüller et al., 2009; Raposo et al., 2009; Shtyrov et al., 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 verbs (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 et al., 2005; Tomasino et al., 2008; Vukovic et al., 2017) or having them carry out a concurrent motor task (e.g., Boulenger et al., 2006; Yee et al., 2013). A causal role can be inferred if taxing parts of our motor system that map onto specific body parts (e.g., the arms) selectively interferes with processing of action verbs that refer to arm-related actions (e.g., clap), but not with words that relate to other body parts (e.g., kick).

Interference is also a common method in studies on working memory, where interactions between a concurrent task (e.g., motor movements) and working memory performance provide 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 et al., 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 et al., 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 et al., 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 et al., 2018; Pecher, 2013; Pecher et al., 2013; Quak et al., 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 et al., 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

[1] The accepted Stage 1 manuscript and protocol of this Registered Report was registered on the Open Science Framework (OSF) and can be found at https://osf.io/v5zm7.

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 verbs 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 6-s memorization phase, they were asked 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 (Fig. 1). The results showed a cross–over interaction effect indicating that arm and leg movements led to effector-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 (Fig. 1 inset).

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 verbs 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

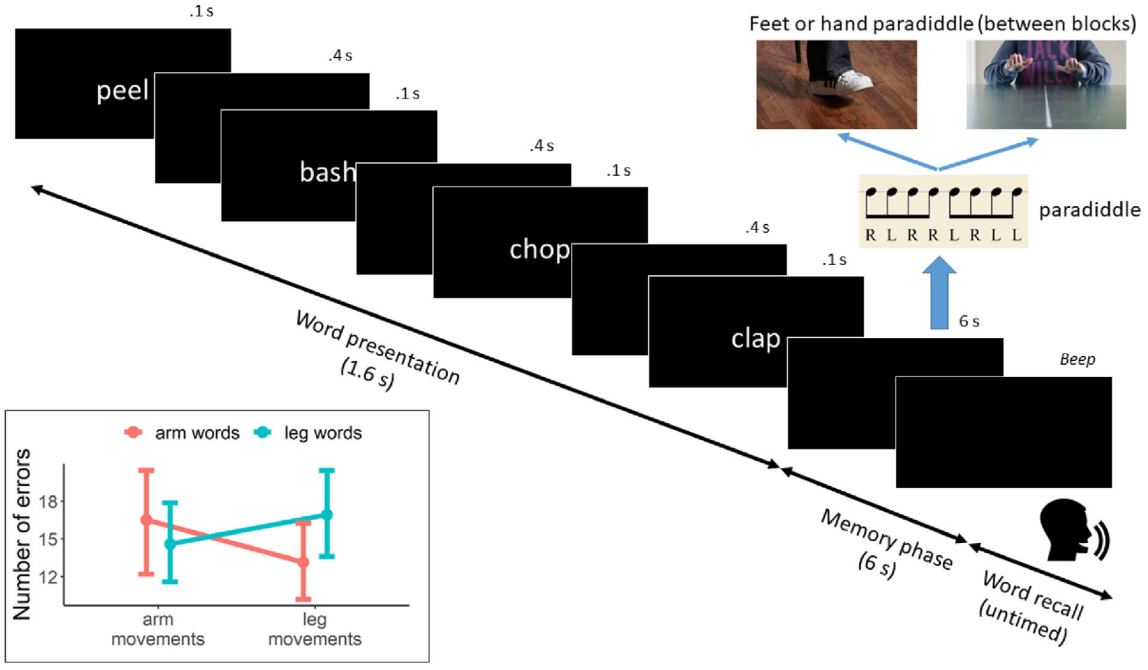

Fig. 1 – Trial structure and experimental design in SP13; inset figure shows original results. In each trial, participants saw a sequence of four different words that were either all arm-related or leg-related (between trials, within blocks). Words were shown for 100 ms with a stimulus onset asynchrony 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 s, either with their hands or with their feet (between blocks, within subjects). After 6 s, 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).

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).[2] 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 et al., 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 $d = .8$ (Cohen, 1988). However, as detailed in Appendix B, we were not able to reproduce this effect size when re-analyzing the original data. Our re-analysis with a more appropriate Bayesian binomial mixed model yields a 95% credible interval for the critical interaction effect of [.05, .24] log-odds; while the interval does not contain zero, the Bayes factor of the alternative against the null was $BF = 1.7$, yielding only anecdotal evidence in favor of the alternative hypothesis (Appendix B). Our simulations equally suggest that the original study was underpowered to detect the very effect they reported (see section "Sample size rationale" and Appendix C). 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 et al., 2011, p. 1363). SP13 might be an example of such a study, thus warranting an appropriately powered replication.

Finally, SP13 analyzed 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 .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 analyze the data with mixed logistic regression, as the probability model underlying this analysis is well suited for error count data (see Jaeger, 2008). Additionally, subject- and item-level variability can simultaneously be modelled, leading to improved inferences about population-level effects (Baayen et al., 2008; Gelman & Hill, 2007).

In sum, SP13 is a study of high theoretical relevance because it supports a causal role of the motor system in keeping action verbs in working memory. However, there are also good reasons to attempt a replication of their result: The direction of the effect was not predicted; a later similar study by the same authors did not yield equally convincing results; the sample size was of only 15 participants; and the statistical analyses were inappropriate for the data.

Our aim was to run a direct replication of SP13, preregistering all aspects of data collection and data analysis, and introducing only minimal changes to the original design (detailed below). We sought to replicate the finding that executing arm or leg movements selectively impairs working memory for arm- and leg-related action verbs, respectively. This constitutes a strong test of the claim that the sensorimotor system is "*necessary* for action-word memory" (SP13, p. 227, emphasis in original). To plan for compelling evidence, we adopted 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 $N = 108$ (over seven times the original), with step sizes of 12 participants. We defined a clear stopping rule for data collection based on a predetermined threshold as to what constitutes evidence for or against the alternative hypothesis using Bayes factors (BFs) (Dienes, 2014; Verhagen & Wagenmakers, 2014). The expected statistical power of our study was high (>90%) based on a simulation-based design analysis (see Sample size rationale below).

Given the mixed evidence for interference effects in the embodiment literature and the fact that strong claims have been made based on small-sample studies, the outcome of our replication is an important reference point in the field. First, this replication had a maximal sample size over seven times that of the original and four times the median sample size of the 33 experiments in the 12 studies we reviewed on working memory and motor interference (median: 27; range: 16–52; see Appendix H). Second, we adopted an appropriate statistical tool to analyze recall data (logistic mixed regressions), thereby increasing the sensitivity of our estimates without

[2] 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 the results in SP13 so compelling.

inflating false positive rates (Jaeger, 2008). Third, our fully pre-registered approach reduced the possibility of conscious or unconscious bias by curtailing researcher degrees of freedom (Simmons et al., 2011). Finally, the Bayesian analysis means that the weight of the evidence, even if inconclusive with respect to the hypothesis, is informative, both in terms of quantifying the results of the replication attempt (Verhagen & Wagenmakers, 2014) and in providing a credible interval for the magnitude of the effect of interest through the posterior distribution (Kruschke, 2010). In sum, if the effect replicated, the present study would provide a template for other researchers in the field for how to move forward carrying out studies that adhere to the standards of reproducible science (Munafò et al., 2017). If the effect did not replicate or if the results remain inconclusive with a sample of over 100 participants, it should lead to a re-evaluation of our theories or at least of the predictions that provide strong tests of these theories (Platt, 1964).

2. Method

Fig. 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 followed 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. *Sample size rationale*

We adopted a prospective Bayes factor design analysis to plan sample size (BFDA, Schönbrodt & Wagenmakers, 2018). In contrast to *p* value-based inference, using BFs 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 (H1) that there is an effect, b) support the null hypothesis (H0) 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 (N_{min}), 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 (N_{max}) at which for practical considerations data collection stops, irrespective of the degree of evidence reached.

We used the Monte Carlo method for our design analysis (see Johnson et al., 2015). Here we outline the general approach and synthesize the outcome of the simulations; see Appendix C for details. We generated a large number of data sets with parameter values taken from our re-analysis of the original data of SP13 and our own pilot data (pilot data was used for parameters that could not be estimated from the original).[3] All simulated data sets consisted of trial-level data with 104 items per participant, as in our actual design. Each data set was randomly generated under a probabilistic binomial (Bernoulli) hierarchical model in which the log-odds of producing an error were a function of the population-level (fixed) effects predictors Interference Movement (arm movements *vs* leg movements), Word Type (arm-related *vs* leg-related words), and their interaction. In addition, random effects variance was added by participants (for intercepts and all the fixed effects and interaction slopes) and items (for intercepts and slopes for Interference Movement). The simulations crossed the following factors:

- Participant sample size: $N = 15, 60, 108$; that is, the original sample size, N_{min}, and N_{max}, respectively.
- Simulation type: Type 1 (critical population-level effect set to zero), type 2 (critical population-level effect sampled from the model of the original data).

Each simulated data set was analyzed with two binomial mixed models using *lme4* (Bates et al., 2015), one that contained the critical interaction (Interference Movement-by-Word Type) and one that did not. A Bayes factor was then computed for the alternative hypothesis that the interaction is different from zero (H_{10}), using the Bayesian Information Criterion (BIC) approximation of the Bayes factor (Wagenmakers, 2007).[4] Following *Cortex* guidelines, we set the threshold for accepting the alternative over the null hypothesis (or vice versa) at a Bayes factor of 6 ($BF_{10} \geq 6$ or $BF_{01} \geq 6$). This allows us to evaluate Type 1 and 2 error rates under our current design.

Fig. 2 summarizes the results of the simulations (10,000 simulations for each combination of sample size and simulation type).[5] The left panel (type 1 simulations) represents cases in which the population-level effect of the critical interaction is set to zero. It shows the proportion of cases in which we would either correctly accept the null (H0), remain undecided, or incorrectly accept the alternative hypothesis (H1). The latter case (type 1 errors) almost never occurred, suggesting that false positive rates are extremely low given our design and analysis method. Even the rate of inconclusive evidence was low (<1.5%) for all three sample sizes.

The right panel in Fig. 2 (type 2 simulations, a Bayesian version of a power analysis) represents cases in which the effect really exists and is of a magnitude comparable to that in the original. Here, the sample size matters. For $N = 15$ (the original sample size), our inferences would be very poor: We would correctly accept H1 only 17% of the time; the data would

[3] The original data from SP13 are available at https://zenodo.org/record/3402035#.XZjAJkb7RaQ.

[4] The BIC approximation is computationally much cheaper than the fully Bayesian approach using bridge sampling that we will adopt for our actual analyses. Our simulations took about a week running on a computer cluster but would have taken several months had we used bridge sampling. For a comparison of different methods to compute BFs, see (Lindeløv, 2018).

[5] We report only simulations for which the models converged; see Appendix C in the Research Compendium for convergence failure rate.

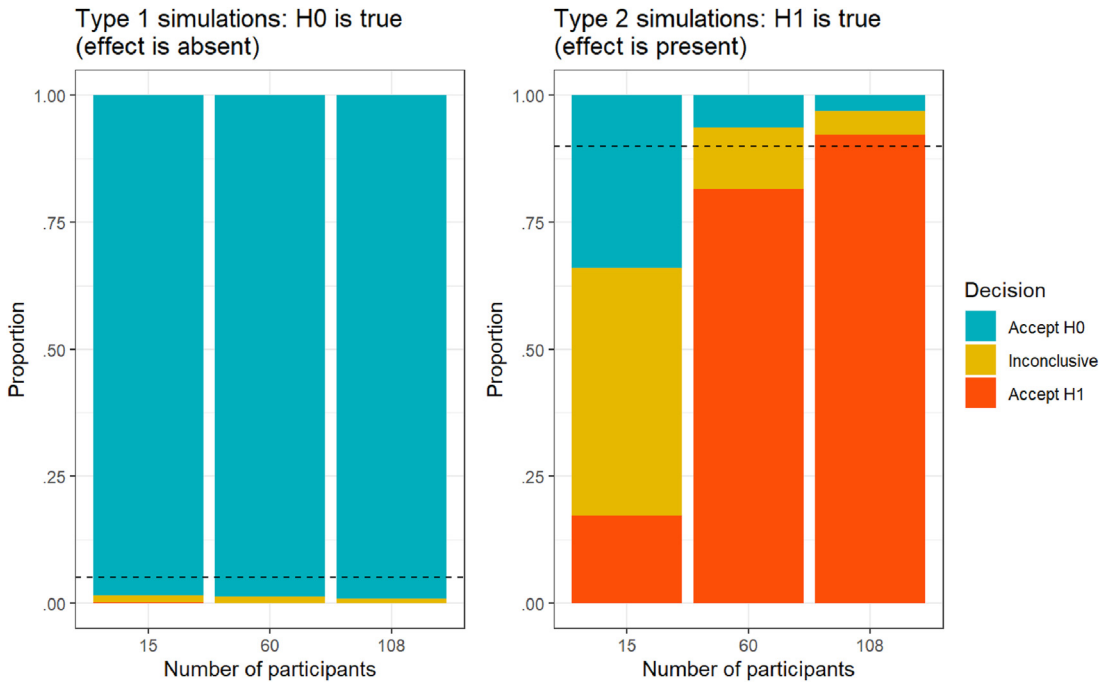

Fig. 2 – Summary of Bayes factor design analysis. For each simulated data set, the decision could be to either accept the H0 (if BF ≤ 1/6, blue fill), remain undecided (if 1/6 < BF < 6, yellow fill) or accept the H1 (BF ≥ 6, red fill). The plots show the proportion of decisions per sample size (*x*-axis) and simulation type (left and right panel). In type 1 simulations (left panel) the critical population-level effect is absent: H0 is true and accepting it is the correct decision. The dashed line at 5% shows the conventionally accepted rate of mistakenly rejecting H0. In type 2 simulations (right panel) the critical population-level effect is present: H1 is true and accepting it is the correct decision. The dashed line at 90% shows the minimal power required by this journal. Each bar is based on at least 10,000 simulations.

be inconclusive in 49% of cases; and we would incorrectly accept the H0 on 34% of occasions. In contrast, for $N_{min} = 60$ we would correctly accept H1 82% of the time and incorrectly accept H0 only 6% of the time (with 12% inconclusive evidence). Finally, for $N_{max} = 108$, we would correctly accept H1 92% of the time, incorrectly accept H0 in 3% of cases, and remain undecided in 5% of studies.

We emphasize that the only difference between the type 1 and type 2 simulations is that the former set the critical population-level interaction effect to zero, while for the latter it is based on our re-analysis of the original data (sampled from a normal distribution with mean equal to the mean estimate and SD equal to the SEM). All other sources of variance (fixed and random effects) are the same in both simulation types (see Appendix C).

2.2. *Participants*

2.2.1. *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.

2.2.2. *Our replication*

Our study was conducted in Sweden and we recruited native speakers of Swedish in the same age range as the original (18–30). We adopted a sequential Bayes factor design (Schönbrodt & Wagenmakers, 2018) with a minimum sample size of 60 and a maximum sample of size 108 participants with step sizes of 12 participants. Participants excluded from the statistical analysis due to pre-specified exclusion criteria (see below) were replaced by new participants; the number of exclusions is reported below. The exact sampling plan was as follows:

1. Collect data from $N_{min} = 60$ participants.
2. Compute the BF with a weakly informative prior (see Analyses below).
3. If $BF_{10} \geq 6$ or $BF_{01} \geq 6$, stop data collection and report results. Else:
4. If $N < N_{max} = 108$, collect another batch of 12 participants and go to step 2. Else:
5. If we reach $N_{max} = 108$, stop data collection, compute BFs and report results.

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

2.2.3. *Notes after data collection (including any deviations)*

There were two deviations from our initial plan for which we here provide the relevant context. First, our final sample size is $N = 77$, which deviates from our plan to first collect data from 60 participants and then add batches of 12 participants if the results were inconclusive. Upon reaching our first milestone of $N = 60$ participants, the data was not immediately ready to run our pre-specified analysis. We had to first run the inter-rater reliability analysis, where 5% of the data had to be transcribed by two independent native speakers of Swedish (as pre-registered). Once inter-reliability was verified (see Results below), the rest of the data needed to be transcribed before we could analyze it. These steps took in the order of four weeks, which was more time than we had anticipated. Since we could not be sure if the outcome of the analysis would be decisive and we had hired a research assistant specifically for this project, we decided to continue running participants in the meantime. Once we had run the main analysis on our first 60 participants (as pre-registered) and the results were conclusive in favour of the null (see analysis report in Research Compendium), we stopped data collection. At this point we had reached a total of $N = 77$ participants. Below we report the results with the full data set because this constitutes the most robust empirical evidence. The results with $N = 60$ lead only to very marginal differences compared to the ones reported below and do not affect the overall interpretation. The analyses with $N = 60$ are reported in the Research Compendium.

The second deviation concerns an increase in the age limit for participants from 18-30 years to 18–40 years, five months into data collection. We provide the relevant context: Data collection started end of October 2020 but, due to ongoing Covid19 restrictions, proceeded slower than anticipated (just about 60 participants after more than four months of data collection by a dedicated research assistant). To increase recruitment rate, we opted for increasing the age criterion for participants from 18-30 years to 18–40 years. This change received editorial approval from *Cortex* on 3 May 2021, as there was common agreement that there was not a theoretical argument for excluding people in the 31–40 years age range and that the increase in risk of bias was negligible (see Appendix K). No further changes were made to our initial plan.

Nine participants were excluded and replaced (before their data was analyzed): two because of technical failure; five because of history of neurological or psychiatric diagnoses (dyslexia, autism, depression, psychosis)[6]; one because they were not born in Sweden; one because they did the experiment twice (second session was excluded). Participants' mean age was 25.4 years (SD = 4.6; range: 18–40). Among them, 46 were female, 27 male, and four indicated "other" or did not disclose their gender. The average speed for hand and foot tapping of paradiddles was 210 (S.E.M. = 6.7) and 189 beats per minute (S.E.M. = 4.9), respectively. All participants indicated they were right-handed and their score on the handedness questionnaire confirmed this (mean = 55.7, SD = 16.7, range: 15–100; positive scores indicate right-handedness while negative ones indicate left-handedness). All participants included in the analysis satisfied the selection criteria specified above.

2.3. *Materials*

2.3.1. *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 = .14]; leg words = 1.92 [.12]) and leg-relatedness (arm words: 2.28 [.13]; leg words = 5.58 [.22]), as assessed by semantic ratings (the scale is not reported in SP13).

2.3.2. *Our replication*

To increase statistical power (Brysbaert & Stevens, 2018), we increased 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).[7] Crucially, our two lists also differed significantly on arm-relatedness (arm words: 6.59 [SE = .03]; leg words = 1.80 [.07]) and leg-relatedness (arm words: 1.34 [.03]; leg words = 6.46 [.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.

2.4. *Procedure*

2.4.1. *Original study*

The basic procedure is shown in Fig. 1. Each trial began with a fixation point shown in the center of the screen for 3 s. 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 s memory phase during

[6] These participants had clinical diagnoses that met our exclusion criteria but did not report this until their exit questionnaire.

[7] 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).

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.

Feature	Arm words		Leg words		*p* value (*t*-test)
	Mean	SE	Mean	SE	
Number of letters	5.13	.13	5.37	.18	.3
Number of phonemes	4.69	.1	5.02	.16	.1
Word log frequency	2.56	.09	2.28	.13	.1
Lemma log frequency	2.79	.09	2.62	.13	.3
Bigram log frequency	6.02	.04	6.03	.05	.8
Trigram log frequency	4.82	.07	4.84	.07	.8
Grammatical ambiguity	.2	.02	.16	.02	.2
Valence	3.67	.1	3.79	.11	.4
Arousal	2.49	.09	2.32	.09	.2
Imageability	5.54	.06	5.33	.1	.1
Arm-relatedness	6.59	.03	1.8	.07	<.001
Leg-relatedness	1.34	.03	6.46	.08	<.001

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 orally recall the four words in the order they had encountered them. Participant responses were audio-recorded for later transcription. 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 s. Each participant's hand/foot frequency threshold was defined as the highest pace at which they could maintain error-free performance for 20 s. 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 s memory phase, and an articulatory condition, in which participants had to repeat the syllable *bla* throughout the memory phase. The latter was not included in our replication as there is no theoretical reason to assume that the embodiment effect depends on participants also performing the articulatory suppression condition.

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 could 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 would amount to 36 trials (i.e., 144 words with four words per trial). We checked with the authors who clarified that the former figure (24 trials) was the correct one, noting that "48 words from each category [of which 36 were unique words] were shown in each block. Twelve words per category, randomly selected, were repeated once in each block" (Z. Shebani, personal communication, April 1, 2018).

2.4.2. *Our replication*

Our replication followed the procedure reported in SP13 with the following exceptions. First, we included the two critical conditions (hand-movement and foot-movement) and the control condition but omitted the articulatory condition. Strictly speaking, only the hand- and foot-movement conditions are relevant to the tested hypothesis, as made clear in SP13, who consistently refer to these as the "critical conditions […] directly addressing the main hypothesis motivating this study" (p. 225–226). We kept, however, the control condition to allow for data and quality checks, such as assessing how many errors people made and whether errors varied systematically between arm-related and leg-related words in the absence of interference (see Quality checks below). The order of the conditions was counterbalanced across participants.

Second, we assumed that repeating a random subset of 12 out of 36 unique words per category and block (as in the original) was not critical to the obtained result and thus opted for a more standard design in which each word is shown once per block. Since we have a larger set of stimuli (52 words), this increased the trials per block to 26 compared to SP13's 24.

Third, we used three (rather than two) random lists grouping the same 104 stimuli words into different 4-word items (each quadruple always consists of either arm or leg words).[8] Each participant saw each list once (one per block), with the assignment of lists to block type counterbalanced across participants (Appendix F). The specific order in which the items of a list were 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.

Fourth, we implemented a set-up that allowed us to monitor performance on the paradiddle tasks. Two digital drum pads (model: Alesis Samplepad 4) recorded participant hand/foot tapping during the interference conditions. Each device sends MIDI information that is logged together with the output of the experiment and can then be mapped as left/right taps from the corresponding effector, linked to a time stamp. This information was used to exclude participants who systematically failed to carry out the rhythmic task (see Exclusion

[8] 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.

criteria below). 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). We agree that the number of rhythmic errors is not the focus of interest but wanted to ensure that participants were engaged in the motor task, as this is a prerequisite to test the critical hypothesis. Before debriefing at the end of the experiment, we asked participants if they had an idea of what exactly the study was about.

2.5. Data exclusion criteria

At the trial level we applied two types of exclusion criteria:

1) We excluded trials in which the participant started the oral recall before the beep, that is, if the word onset fell before the end of the 6 s memory phase (see Fig. 1).
2) In the two interference conditions, we excluded trials in which participants failed to execute the interference task, which we define as starting the paradiddle later than 3 s into the memory phase (that is, if the first tap is registered later than 3 s after the offset of the fourth word in the trial).

At the participant level, we excluded participants for whom either of the above criteria or technical failure (e.g., recording not working) resulted in excluding more than 30% of the trials across blocks or more than 50% of trials in a single block.[9] All exclusions took place before the recall data was coded and analyzed. Excluded participants were replaced.

2.6. Quality checks

As a quality check we verified that there were 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 analyzed 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, 2020 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 = .39 log-odds, SE = .032, $p < .001$; analyzed 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 tested 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 included a by-participant random intercept and random slope for word position within trial, as well as a random intercept by verb.[10] For this analysis, we used the R package *lme4* (Bates et al., 2015).

Finally, we analyzed performance on arm and leg-related verbs in the control condition to establish if there were category differences in recall independently of the interference task. This serves as an additional check of our Swedish stimuli, which is however orthogonal to the crucial test of the 2 × 2 interaction.

3. Data coding and analyses

3.1. Data coding

We adopt a binary coding for the oral recall data: For each word within a 4-word memory trial, the dependent variable is 1 if there is a memory error (verb not recalled) and 0 if there is no error (verb correctly recalled). Thus, there are 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, on theoretical grounds, we are not aware of any embodiment proposal that predicts interference effects would specifically result in sequencing errors for effector-congruent words. Importantly, and as just 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). Second, 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 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

[9] We take the prediction of the embodiment hypothesis to be that engaging in a complex motor task should lead to effector-specific interference. We will therefore not exclude trials based on imprecise execution of the paradiddle, as interference could in principle be bidirectional, from movements to words and from words to movements (see García & Ibáñez, 2016); if so, removing trials with execution errors would potentially remove critical trials where the hypothesized interference is taking place. Our exclusion criteria focus on participants *using motor skills* to execute the interference task, even if their execution is imperfect.

[10] Model formula 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).

be coded as an omission and an addition (two errors). For want of a principled protocol that can be implemented in a machine, we prefer to adopt our more transparent coding scheme. Third, 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), the same as if the response had been *peel-potato-garden-I don't know* (two replacement errors and an omission). Intuitively, the former response is superior to the latter in recall, but this would not be captured by the coding. Finally, from a measurement-theoretic viewpoint, our coding scheme allows for improved inference on population-level effect estimates by letting us model participant and item variability as random effects. 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.[11]

3.2. Inter-rater reliability

Initially, a randomly selected 5% of observations from the first 60 participants (i.e., 624 data points) were transcribed and coded independently by two raters who were native speakers of Swedish. If the inter-rater agreement was ≥95%, each of the raters would proceed to code separate subsets of the complete data set. If inter-rater agreement was <95%, disagreements would be inspected and resolved through discussion, so that coding criteria become shared among raters. Then a separate 5% sample of the data would be coded, and the procedure repeated until inter-rater agreement reached ≥95%. The number of rounds needed to reach threshold and the agreement rate at each round is reported below.

3.3. Analytic approach: bayesian logistic mixed effects regression

We analyzed 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 followed the guidelines in Barr et al. (2013): We started 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 simplified this random effect structure in the principled way outlined in Appendix G. Additionally we included the following nuisance variables as fixed effect predictors in the model (centered 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 standardized predictors and then set a weakly informative prior for all coefficients: a Normal distribution centered 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 used 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 report mean estimates and modes, standard errors, and 95% credible intervals for all fixed effects model parameters. The dataset and analysis script are openly shared in our Research Compendium.

3.4. Stopping rule and assessing the outcome of the replication with bayes factors

To decide when to stop data collection (see Participants) and to make a decision as to whether our replication successfully detects the effect reported in SP13 or fails to do so, we used 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 computed 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, σ = 2)), while BF2 uses as prior a normal distribution based on the posterior estimates of the model fitted to the original data in SP13.

[11] 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.

Our decision as to when to stop data collection (see Participants) was based on the calculation of BF1 only. Once data collection stopped (either because BF1≥6 in favor of one of the competing hypotheses or because we have reached $N_{max} = 108$) we computed BF2.

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

4. Results

4.1. *Main result (N = 77): very strong evidence against the hypothesized effect*

Our main result is based on our final sample of N = 77 (see rationale for this final N in section 2.2.3 – 'Notes after data collection').[12] A visualization of the raw data averaged by participant is shown in Fig. 3. There is little visual indication of the hypothesized selective semantic interference effect, which should have manifested itself as a cross–over interaction (see inset in Fig. 1).

Following our pre-registered analysis pipe-line, we computed the Bayes factor in favor of an effector-specific interference effect. To do this, we ran two almost identical Bayesian logistic regression models, differing only in the presence (full model) or absence (null model) of the population-level critical interaction effect. Hence, only the full model could statistically capture a result in which participants' arm movements selectively interfered with memory for arm-related words while leg movements interfered with leg-related words (the hypothesized effect). By comparing this full model to the null model and computing the Bayes factor in favor of either, we are effectively answering the question how much evidence there is for the effect of interest.

The models followed the exact specification in our pre-registration, including all priors for model parameters (see Appendix E). There was no need to simplify the random effect structure. Thus the final formula for the full model in R was: "error ~1 + interference_type * word_type + trial_in_experiment_z + word_position_in_trial_z + preceding_error_z + (1 + interference_type * word_type | subject) + (1 + interference_type | verb)", where *error* is the binary dependent variable (1 if a word is not correctly recalled, 0 if it is) and z means that the predictor is scaled. The null model was identical except that the population-level interaction between interference type and word type was removed. For further details, see the Research Compendium.

The Bayes factor in favor of the null model was $BF_{01} = 91$, that is, we found very strong evidence in favor of the null hypothesis that no effector-specific interference is present (see Table 2). Fig. 4 shows the model estimates for the full model. The 95% credible interval for the critical interference type by word-type interaction term contains 0, which shows there is no statistical support for this effect.

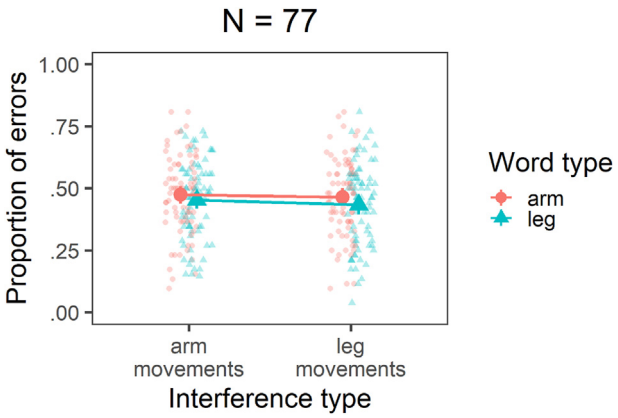

Fig. 3 – Raw data averaged by participant. Each participant contributes two circles (representing the proportion of errors remembering arm-related words when either moving their arms or legs) and two triangles (analogously representing the number of errors remembering leg-related words). The lines are almost parallel, indicating scarce support for the hypothesized interaction effect. Error bars show 95% confidence intervals of by-participant means (non-parametric bootstrap).

Table 2 – Heuristic classification scheme for the interpretation of Bayes factors BF_{10} (adjusted from Schönbrodt & Wagenmakers, 2018). The same scheme is 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

4.2. *Secondary analysis (replication bayes factor): extreme evidence against the hypothesized effect*

Our secondary pre-registered analysis computed a replication Bayes factor (Verhagen & Wagenmakers, 2014). Here we address the question if the "effect from the replication attempt [is] comparable to what was found before, or [is] absent?" (Verhagen & Wagenmakers, 2014). This approach can be understood as a more lenient analysis, as it aligns the Bayesian priors with the outcome of the original study, rather than using priors that assume no knowledge at all. It captures the notion that we have some prior expectations about the effect size (based on the original study to be replicated), and we can ask whether the observed outcome would be expected taking this previous evidence into account. To make an

[12] We report the results for the analysis with the first 60 participants in the Research Compendium. Crucially, the results are very similar and support the same interpretations.

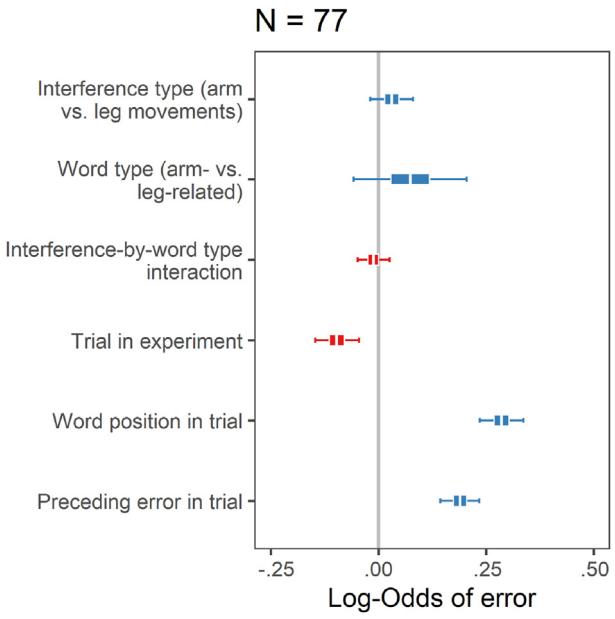

Fig. 4 – Model estimates from the full model, which modelled the hypothesized interaction effect at the population level (Block type-by-word type interaction). The boxplots display high density intervals from the *brms* model (50% probability for the boxes, 95% for the whiskers, and the median for the point estimate; the color of the boxplot depends on the sign of the median; plotted with *sjplot::plot_model*). The 95% credible interval for the critical interaction contains 0, which shows there is no statistical support for this effect.

analogy, it is akin to a situation in which your friend tells you a certain movie is very good (or very bad); when you watch it, you ask yourself not simply if it is a very good (or very bad) movie, but rather whether your friend was justified in describing it as such. Your friend's description becomes the anchoring point of your own evaluation. Here, our re-analysis of the previous data goes in as the prior for the effect of interest.

The replication Bayes factor was $BF_{01} = 221$. That is, we found extreme evidence for the absence of an effect. It may seem surprising at first that the evidence is even *stronger* in this analysis that is supposedly more lenient. The reason is this: Our main analysis assumed no prior knowledge for the critical effect, and thus it was non-directional; in other words, the prior was centered around zero, with tails symmetrically going in either direction (the expected interference effect or a non-expected facilitation effect). In contrast, the replication Bayes factor assumed that the interaction effect went in a specific direction, namely that it would manifest itself as interference (as in the original). Yet, our results show a very slight numerical tendency in the *opposite* direction (towards facilitation, even if it is practically zero). This explains why the outcome is less compatible with the results from the original than it is with a simple "ignorant" prior that assumes no specific knowledge at all (our main analysis).

4.3. *Data exclusion, quality checks and other control analyses*

4.3.1. *Data exclusion*

Following our pre-registered data exclusion criteria based on paradiddle performance (section 2.5), we excluded 66 individual trials, leading to the loss of 264 data points (4 data points per trial) out of a total of 24,024 data points (11%). No entire participants had to be excluded based on our pre-defined criteria.[13]

4.3.2. *Inter-rater reliability in data coding*

We assessed inter-rater reliability in the coding of the data (i.e., in the transcription of participant responses) by having 5% randomly selected observations from the first 60 participants (624 data points) transcribed independently by two raters who were native speakers of Swedish. Both raters received the same transcription guidelines (Appendix I). There was a 98.2% agreement rate between raters, which exceeded our pre-defined threshold of 95%.[14] One of the two raters thus proceeded to transcribe the rest of the data following the same guidelines.

4.3.3. *Absence of floor or ceiling effects*

As a pre-registered quality check, we verified that there were no ceiling or floor effects (0% or 100% of errors) in any of the four experimental cells. As shown visually in Fig. 3 and detailed in Table 3, there were no ceiling or floor effects for participant averages in any of the four cells.

4.3.4. *Positive control*

As our pre-registered positive control, we verified if there was an effect of serial position on word recall, a very robust finding in the working memory literature (see Popov & Reder, 2020). We expected more errors for later than for earlier words in a quadruple. This is an outcome-neutral criterion because it is orthogonal to our main effect of interest. As expected, we found an effect of serial position on recall error, with later words being more likely to lead to recall error. In our preregistered analysis conducted using the *lme4* package we found a significant effect of serial position: estimate = .29, *SE* = .05;

Table 3 – Descriptive statistics for the proportion of errors per experimental cell, averaged by participant.

Experimental cell	M	SD	Range	N
Arm movements, arm words	.47	.16	.1–.81	77
Arm movements, leg words	.45	.16	.15–.73	77
Leg movements, arm words	.47	.16	.1–.81	77
Leg movements, leg words	.43	.17	.04–.81	77

[13] See Research Compendium (file "analysis/trial_exclusion.html") for a report and the corresponding script documenting how the trials were excluded.

[14] See Research Compendium (file "analysis/interrater_agreement.R") for a script documenting the interrater agreement analysis.

Wald's z = 5.77, $p < .001$. Thus, our positive control was verified.[15]

4.3.5. *Comparison to control condition (no interference)*
Finally, we analyzed performance in the control condition, in which participants did not engage in any concurrent task, but simply had to keep the words in memory during a silence period of 6 s. Again, while not central to our hypothesis (which is tested by the critical interaction), this served to verify that our stimuli were not strongly biased such that either arm- or leg-related words were easier to remember. As visually shown in Fig. 5, there was a slight numerical advantage for memory of leg-related words in all conditions. We ran an identical model as the full model reported under the main analysis, but this time including the control condition. Interference type was contrast coded (2 contrasts: arm vs. control, leg vs. control).[16] The estimates were very similar to the main model reported above (see Fig. 4). There was no difference in the likelihood of an error between arm- and leg-related words (estimate = .07; 95% CI = [−.06, .21]). There was a sizeable increase in the likelihood of errors in both interference conditions compared to control (arm vs. control estimate = .36; 95% CI = [.29, .42]; leg vs. control estimate = .30; 95% CI = [.23, .37]). Crucially, there were no interactions between word type and interference type. Since the control condition was included in this analysis, there were potentially two such interactions: First, the effect of verb type (arm *vs* leg verbs) did not differ between the arm movements and control conditions (estimate = −.01; 95% CI = [−.06, .03]); second, the effect of verb type did not differ between the leg movements and control conditions (estimate = .01; 95% CI = [−.04, .06]).

In summary, arm-related and leg-related items were of similar difficulty, both types of interference conditions (arm and leg movements) were harder than the control condition (no movements), but the errors on arm- and leg-related words did not depend on interference type.

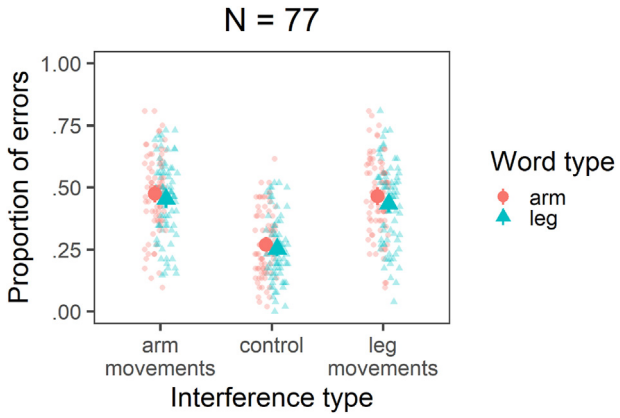

Fig. 5 – Raw data averaged by participant, including the control condition. For details, see Fig. 3.

5. Discussion

We have found very strong evidence that the motor system is *not* needed to keep action verbs in working memory. We conducted a fully pre-registered direct replication of an interference study that is credited by critics of the embodiment hypothesis to "provide the strongest evidence to date for the view that motor simulations support short-term memory" (Zeelenberg & Pecher, 2016, p. 183). The original study found that carrying out a complex rhythmic pattern with the arms selectively interfered with memory for arm-related action verbs, while carrying out the same rhythmic pattern with the feet selectively interfered with memory for leg-related action verbs (Shebani & Pulvermüller, 2013). We carefully replicated the methods of the original, improved the analysis, and increased the sample size to ensure we had enough power to detect the effect, with a final sample size over five times that of the original. Yet, we did not find any sign of the predicted effector-specific interference. In our pre-registered secondary analysis (replication Bayes factor), we found extreme evidence that our results are *not* compatible with the original findings but rather support the null hypothesis. We stress that our Bayesian analysis means we have not merely found absence of evidence, but rather *very strong evidence of absence* of the effect reported in the original study.

At a theoretical level, our findings provide very strong evidence that an involvement of the sensorimotor system is not necessary to keep action verbs in memory. The semantic embodiment hypothesis has been hotly contested, dividing scholars on its theoretical merit. By far, the strongest and most controversial version of semantic embodiment asserts that sensorimotor involvement is necessary and automatic in high-level cognitive processing (for proponents, see Barsalou, 2008; Pulvermüller, 2005; Pulvermüller & Fadiga, 2010; for opponents, see Mahon & Caramazza, 2008; Zeelenberg & Pecher, 2016). The present study provides the most robust demonstration to date against the view that motor simulation plays a functional role in short-term memory of action verbs. In short, we can engage in actual motion with our legs and feet without interfering with our ability to keep the word "run" (as opposed to "clap") in short-term memory.

5.1. *Why did the original finding not replicate?*

Before discussing further implications, we need to address the possible causes of the divergence between our study and the original result in SP13. Ours was a direct replication: we did our best to reproduce the exact procedure that was used in the original and checked with the authors that we were not deviating in any meaningful way. The interference task followed the same protocol and the items had been normed for their leg- and arm-relatedness and controlled for other lexical factors. What can explain the different results?

The most obvious difference between our study and the original is that ours was carried out in Swedish, while the

[15] We report the pre-registered analysis using *lme4*. Note that a very similar estimate can be read off our full Bayesian model (fit with *brms*), see Fig. 4: estimate = .29, 95% credible interval = [.24, .34]. In other words, the success of our positive control seems very robust.

[16] For details of this analysis, see the "analysis/analysis.html" report in our Research Compendium.

original was in English. We do not think that this is a determining factor for the results, since the hypothesis of embodied semantics is not a language-specific theory, but rather one that aims at describing the basic human cognitive and neural architecture. If motor simulations play a functional role in processing action verbs, they should do so in any language, not just in English.

There is, however, one interpretation under which a language difference could be expected: In the absence of any linguistic context, an English verb like *kick* can always be interpreted as an infinitive ("to kick") or an imperative ("kick!"). Since the original study used bare infinitives ("kick"), it could be that participants were interpreting them as instructions to perform the relevant actions and thus were (unconsciously) activating their motor cortex to initiate those actions, which in turn led to interference. In Swedish, most, but not all our stimuli words conformed to that pattern: 85% of the verbs we used were both imperatives and infinitives, as in English (e.g., *borsta* = brush); but 15% were unambiguously in the infinitive form (e.g., *gripa* = to seize/catch, whose imperative form is *grip*). Thus, it could be that English participants were interpreting the words as invitations to perform actions (i.e., as imperatives) while Swedish participants were not.[17] First, it should be noted that, even if this *were* the true cause of our failure to replicate, our main conclusion would still hold. If the effect were to hinge on the imperative form of the verb, then motor simulations would by definition not be *necessary* to keep action verbs in memory; they would depend on people actively thinking about performing the actions. However, given that our study was well-powered, we still tested this possibility by running a post-hoc analysis identical to our main analysis but using only those items that could be interpreted in the imperative form (as in the original study). The evidence against an effect for this subset of verbs was as strong as in the main analysis ($BF_{01} = 92$ in favor of the null hypothesis). Thus, there is neither a theoretical nor an empirical ground to believe that the language difference between our study and the original had any bearing on our main results.[18]

We think the more likely explanation for the non-replication is that the effect reported in the original was a false positive. False positives can occur in any empirical study simply due to chance. Extensive empirical evidence and theoretical arguments in recent years have demonstrated that the number of false positives might be larger than the conventional alpha-values used in our field to reject the null hypothesis (Bishop, 2019; Button et al., 2013; Ioannidis, 2005; Nosek et al., 2018; Open Science Collaboration, 2015; Simmons et al., 2011). All other things being equal, a crucial factor to increase our confidence in empirical results should be the sample size (Button et al., 2013). In that regard, the results from our replication are more trustworthy than the original results.

[17] We thank Julia Misersky and Peter Hagoort for raising this point during a presentation.

[18] We thank the reviewer Andrew D. Wilson for suggesting running this post-hoc analysis. The details can be found in the "analysis/analysis.html" report of the Research Compendium.

5.2. *Implications for the field of embodied cognition: shifting the burden of proof*

How does our finding apply to the broader field of research on the embodiment hypothesis? Of course, our results do not constitute direct evidence against every conceivable functional role of the sensorimotor system in conceptual processing. It is possible that some sensorimotor processes (e.g., visual simulations) play a crucial role while others (e.g., motor simulations) do not. Perhaps motor simulations matter only for object representations (see Davis et al., 2020; Yee et al., 2013), but not for action representations as tested here. It is also conceivable that the null effect we found is restricted to tasks where processing meaning is not required to solve the task, thus showing that sensorimotor processes are not fully automatic nor always necessary. In line with this interpretation, one could adduce that motor interference effects in semantic processing have been reported for tasks that do involve a semantic judgment (animal vs. non-animal judgments in Davis et al., 2020; concrete vs. abstract judgments in Yee et al., 2013). Similarly in the visual domain, a causal role for low-level visual processes has been shown to hold when the task required participants to process visual information, but not when it did not (Ostarek & Huettig, 2017).

The above defensive moves all constitute sensible suggestions but given the mixed findings in the literature and the current study supporting the lack of a functional role, we think the burden of proof needs to shift towards those proposing that a functional role exists. Two requirements need to be satisfied. First, the theory must be motivated as to why embodiment would be split up that way and, second, qualifying statements to the more general hypothesis must be tested properly in follow-up studies. Regarding the latter point, the field would benefit from running critical experiments as pre-registered studies, as we elaborate next.

The field of embodiment research has known a thriving theoretical discussion about what kind of experimental evidence supports which views and there is a fair degree of consensus as to what would constitute strong evidence for a functional role of the sensorimotor system in higher-level cognitive processing (Mahon & Caramazza, 2008; Ostarek & Bottini, 2021; Ostarek & Huettig, 2019; see also Mahon, 2015 and contributions to that special issue). What has been lacking is an equally forceful discussion about what experimental standards need to be upheld for the empirical evidence to be convincing in the first place. We know from the broader field of cognitive psychology what the problem is: small sample sizes and researcher degrees of freedom lead to false findings, threatening the very foundations of theoretical progress (Button et al., 2013; Ioannidis, 2005; Open Science Collaboration, 2015; Simmons et al., 2011; Zwaan et al., 2018). We also know one very effective solution to this problem: high-powered, pre-registered studies (Chambers et al., 2015; Nosek et al., 2018).

We illustrate the importance of pre-registering exactly which analysis will test which hypothesis with an example from Davis et al. (2020). Davis and colleagues conducted two experiments with 200 participants in each, a truly large sample size by any current standards, which was motivated by an

a priori power analysis (Davis et al., 2020). Their first experiment is presented as a conceptual replication in the visual domain of the motor interference effect reported in Yee et al. (2013). Both studies reported reaction times and accuracy data. However, while Yee and colleagues found an interference effect for accuracy, but not reaction times, Davis and colleagues found the exact opposite: an effect for reaction times but not for accuracy. In their overall interpretation, however, each study focused on the measure where the predicted effect was *present* but it did not further discuss why the effect was not reflected in the other measure. Davis et al.'s replication combined a strong design with a substantial sample size, but it would have been more convincing if the critical measure to test the hypothesis had been defined *beforehand* and, ideally, on theory-relevant grounds. Preregistrations require authors to restrict their degrees of freedom (Simmons et al., 2011) and commit to what they think will provide the best test of their hypothesis (Nosek et al., 2018).

While acknowledging that there should always be room for exploratory work, our current argument is that much progress could be made by agreeing on the empirical foundations of the field: the facts. One way to achieve this is by focusing efforts on high-powered pre-registrations of paradigms that warrant strong inference (Platt, 1964). A promising strategy is for researchers who champion opposing theoretical views to agree on an experimental design that will prove one side right and the other wrong and join forces to run such conclusive studies (Hofstee, 1984).

6. Conclusion

In conclusion, we have shown here that involvement of the motor system is not necessary to keep action verbs in short-term memory. The strong evidence we find here against an effect predicted by semantic embodiment, together with the inconclusive designs in much of the past research on the embodiment hypothesis (Ostarek & Bottini, 2021; Ostarek & Huettig, 2019), shifts the burden of proof towards those theoretical views that posit a functional role of the sensorimotor system in high-level conceptual processing. We invite the field of embodiment research to engage in appropriately powered pre-registrations of crucial paradigms. Only this will ensure solid scientific progress.

Author note

In accordance with the Peer Reviewers' Openness Initiative (Morey et al., 2016), all materials and scripts associated with this manuscript are available as a Research Compendium at https://osf.io/mvz3f/(see also list of appendices at the end of this manuscript).

Credit author statement

Guillermo Montero-Melis: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, writing - review and editing, visualization, project administration, funding acquisition.

Jeroen van Paridon: Methodology, software, formal analysis, data curation, writing - review and editing, visualization.

Markus Ostarek: Conceptualization, methodology, writing - review and editing.

Emanuel Bylund: Conceptualization, writing - review and editing, project administration, funding acquisition.

Open practices

The study in this article earned Open Data and Preregistered badges for transparent practices. Materials and data for the study are available at https://osf.io/mvz3f/

Declaration of competing interest

None.

Acknowledgements

For many helpful suggestions and comments, we thank T. Florian Jaeger and the Human Language Processing lab at University of Rochester; Peter Hagoort, Mante Nieuwland, and the Neurobiology of Language department at the Max Planck Institute for Psycholinguistics; Vencislav Popov; and Maryann Tan. Thanks to Phillip Alday for statistical advice. The valuable feedback from Andrew D. Wilson and an anonymous reviewer substantially improved our pre-registration. Thanks to Petrus Isaksson for data collection, and Pia Järnefelt and Margareta Majchrowska for help with practical preparations. This work was supported by the Swedish Research Council [grant 2015-01317 to Emanuel Bylund and 2018-00245 to Guillermo Montero-Melis].

Appendix

Research Compendium: Repository including data, analysis scripts, and appendices

All anonymized data, full analysis scripts to reproduce the results reported here, and other study materials are publicly accessible as a Research Compendium on OSF at https://osf.io/mvz3f/. Please read the README file in that repository for a detailed list of the resources it contains, including a description of the data files. The following appendices are also included in that repository (Appendices A–H were part of our Stage 1 submission; Appendices I–M were added to the Stage 2 submission):

- 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
- Appendix H: Sample size in studies investigating interference effects in working memory
- Appendix I: Transcription guidelines
- Appendix J: Labfolder output
- Appendix K: Editorial approval to change age range to 18–40 years
- Appendix L: Background questionnaire for participants.
- Appendix M: Handedness form filled out by participants (participants filled out the Swedish form [M1], but the English version [M2] is provided for ease of comparison).

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Leukemia & Lymphoma, 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 Medicine and Life, 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). Academic Press. 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 Medicine and Life, 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

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568*(7753). https://doi.org/10.1038/d41586-019-01307-2, 435–435.

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., Grange, J. A., Perugini, M., Spies, J. 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

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex, 66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

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

Davis, C. P., Joergensen, G. H., Boddy, P., Dowling, C., & Yee, E. (2020). Making it harder to "see" meaning: The more you see something, the more its conceptual representation is susceptible to visual interference. *Psychological Science, 31*(5), 505–517. https://doi.org/10.1177/0956797620910748, 0956797620910748.

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), Article 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 and 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 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

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica, 56*(1–3), 93–109. https://doi.org/10.1016/0001-6918(84)90010-6

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 Medicine and Life, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6*(2), 133–142. https://doi.org/10.1111/2041-210X.12306

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001

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

Lindeløv, J. K. (2018, February 3). *How to compute Bayes factors using lm, lmer, BayesFactor, brms, and JAGS/stan/pymc3*. RPubs. https://rpubs.com/lindeloev/358672.

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*, 5. 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., Lewandowsky, S., Morey, C. C., Newman, D. P., Schönbrodt, F. D., Vanpaemel, W., Wagenmakers, E.-J., & 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, Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, Article 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

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. 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. *Journal of Cognition, 4*(1), 5. https://doi.org/10.5334/joc.139

Ostarek, M., & Huettig, F. (2017). A task-dependent causal role for low-level visual processes in spoken word comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(8), 1215–1224. https://doi.org/10.1037/xlm0000375

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science, 28*(6), 593–599. 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. https://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

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review, 127*(1), 1–46. https://doi.org/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*. R Foundation for Statistical Computing. 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., & 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]* (p. 375). https://www.skolverket.se/publikationer?id=2832.

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. *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. https://doi.org/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 Medicine and Life, 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. https://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). Academic Press. 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* (Vol. 41). Behavioral and Brain Sciences. https://doi.org/10.1017/S0140525X17001972