Replication Recipe (Brandt et al., 2013): Pre-Registration of replication of Shebani and Pulvermüller (2013) in *Cortex*

NB: This document complements the manuscript "Does the motor system functionally contribute to keeping words in working memory? A pre-registered replication of Shebani and Pulvermüller (2013, Cortex)", submitted to Cortex *as a Registered Report. Some parts of this appendix are directly taken from our manuscript.*

The Nature of the Effect

1. *Verbal description of the effect I am trying to replicate (optional)*

We try to replicate the finding of effector-specific semantic interference in working memory, as reported in Shebani and Pulvermüller (2013, henceforth SP13). SP13 report that performing a complex rhythmic pattern (a "paradiddle") with either the arms or legs differentially impairs working memory for arm- and leg-related action words. Specifically, participants had to memorize four sequentially shown words per trial during a 6 second memory phase. The words in a trial were either all arm related (arm trials, e.g., *peel-bash-chop-clap*) or leg related (leg trials, e.g., *hop-stomp-limp-skip*). Both types of trials occurred within a block. The critical finding took the form of a cross-over interaction effect between effector (hand vs foot paradiddle) and word type (arm vs leg words). Specifically, participants made more errors remembering arm-related than leg-related words when they performed the hand paradiddle, whereas the opposite was true when they performed the foot paradiddle, in which case they made more errors memorizing leg-related than arm-related words.

2. *It is important to replicate this effect because (optional)*

This result was interpreted as evidence "that body movements and working memory for action-related words share processing resources" and further as supporting "the necessity of sensorimotor areas of the upper and lower extremities for arm- and leg-word processing" (SP13, p. 227). It can thus be seen as evidence supporting embodied semantic theories that claim "language understanding and memory depend on links between linguistic brain systems and sensorimotor domains" (SP13, p. 222).

The theoretical relevance of this study warrants a replication: Taken at face value, the results of this study's interference design constitute strong evidence for embodied meaning representations (Barsalou, 2008; Pulvermüller & Fadiga, 2010), which is a heatedly debated topic in cognitive science (Mahon & Caramazza, 2008; Mahon & Hickok, 2016). Furthermore, this is the only study to our knowledge that indicates that the motor system functionally contributes to action verb semantics in *working memory*. A recent review cites it as one of the studies that "provide the strongest evidence to date for the view that motor simulations support short-term memory" (Zeelenberg & Pecher, 2016, p. 183).

On the one hand, SP13 stands out by its elegant yet simple behavioural design, and its clear-cut pattern of results. On the other hand, with only 15 participants, the original study is also one of those "underpowered studies with perfect results […] that should invite extra scrutiny" (Simmons et al., 2011, p. 1363). More generally, the current replication is motivated by calls for more replication studies in the psychological sciences (Munafò et al., 2017; Zwaan et al., forthcoming)

following a debate on how reproducibility in this discipline was generally low (Anderson et al., 2016; Open Science Collaboration, 2015).

3. *The effect size of the effect I am trying to replicate is (optional)*

In the original paper, the reported Cohen's *d* for the interaction of interest was *d* = 1.25 (SP13, p. 226). In terms of Cohen's (1988) rule of thumb for size effect interpretation, this would be considered an effect that is by some margin larger than "large" (*d* = .8); it would count as a "very large" effect according to Sawilowsky (2009).

There are, however, potential problems with this effect size estimate. We first reanalyzed the original data shared by the authors with the same analysis type, repeated measures ANOVAs.[1] Although we could precisely reproduce their summary statistics and the *F* and *p* values for their ANOVAs, we were unable to reproduce their effect size estimate of Cohen's *d* = 1.25. Second, the suboptimal analysis method (ANOVA) used in SP13 might have led to an inflated effect size estimate. A more appropriate analysis treats the error counts as arising from a binomial distribution and models participants as random effects with corresponding random intercepts and slopes for all the fixed effects. When we re-analyzed the original data with this improved analysis method (a binomial mixed effects model), the 95% credible interval for the estimate of the critical interaction effect does not contain 0 (see next point). However, a Bayes factor analysis of the alternative hypothesis (the estimate is different from zero) against the null hypothesis (the estimate is zero) only led to anecdotical evidence in favour of the alternative (BF_{10}=1.7). The effect size estimates in log-odds is. For details of our re-analysis, see Appendix B in our submission.

4. *The confidence interval of the original effect is (optional)*

Confidence intervals were not reported in the original. In our re-analysis of the data (see point 3), the 95% Bayesian credible interval of the interaction effect in log-odds was [0.05, 0.24].

5. *The sample size of the original effect is (optional)*

N = 15 (SP13, p. 223).

6. *Where was the original study conducted? (e.g., lab, in the field, online) (optional)*

This is not explicitly mentioned in the article, but most likely in the lab judging from the description of the procedure.

7. *What country/region was the original study conducted in? (optional)*

This is not explicitly mentioned in the article, but probably in Cambridge, UK, since a) the first author was affiliated to the Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, UK, and b) ethics permission for the study was obtained from the Cambridge Psychology Research Ethics Committee.

[1] The original data are available at https://github.com/zshebani/LMB/tree/1.0 (DOI: 10.5281/zenodo.3402035).

8. *What kind of sample did the original study use? (e.g. student, Mturk, representative) (optional)*

The authors provide the following description (SP13, p. 223):

> Fifteen monolingual, native speakers of English (8 males) aged 18–30 (mean = 20.4, standard deviation SD = 3.2) took part in the experiment. All reported normal vision and hearing and had no history of neurological or psychiatric illness. All participants were also right-handed with an average laterality quotient (Oldfield, 1971) of 80.5% (SD = 23.6) and reported no immediate left-handed family members (parents or siblings). All gave their informed, written consent prior to their participation and were reimbursed for their time.

9. *Was the original study conducted with paper-and-pencil surveys, on a computer, or something else? (optional)*

The critical task was administered on a computer. Participants' oral responses were recorded and later transcribed and coded (Z. Shebani, personal communication, October 3, 2018).

Designing the Replication Study

10. *Are the original materials for the study available from the author? (optional)*

Yes, the original stimuli (72 English action verbs that are either arm [36 verbs] or leg related [36 verbs]) are provided as an appendix in the original study (SP13, p. 229). However, we will not use the original materials, as our replication will be conducted in Sweden with native speakers of Swedish.

11. *I know that assumptions (e.g., about the meaning of the stimuli) in the original study will also hold in my replication because (optional)*

We will use Swedish verbs in our replication with native speakers of Swedish: 52 arm-related and 52 leg-related verbs. Arm- and leg-relatedness was assessed by semantic ratings on a 7-point scale obtained from 12 Swedish native speakers. The two lists differ significantly on arm-relatedness (arm words: 6.59 [SE=0.03]; leg words = 1.80 [0.07]) and leg-relatedness (arm words: 1.34 [0.03]; leg words = 6.46 [0.08]). We have matched the two lists of verbs 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). 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).

Because the stimuli are controlled along the same variables as in the original, the only difference is that the study will be carried out in a different language (Swedish instead of English). Since there is no theoretical reason for why the effect reported in SP13 should be language-specific, all the assumptions about the meaning of the stimuli will also hold in our replication.

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

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

12. Location of the experimenter during data collection (optional)

The experimenter will be in the room during data collection, since he/she will have to show the participants how to carry out the rhythmic patterns with hands/feet, and monitor that participants carry out the interference task correctly.

13. Experimenter knowledge of participant experimental condition (optional)

All critical manipulations are within participants. The effector manipulation (arm vs leg paradiddle) is between blocks; the experimenter knows if a participant is in the arm- or leg-interference condition, since the experimenter has to show and practice the paradiddle with the participant and the correct realization of the paradiddle needs to be monitored by the experimenter. The word type manipulation (arm- vs leg-related words) is between trial (within each block); the experimenter does not know in advance which word type a given trial belongs to, as trials are randomly ordered.

14. Experimenter knowledge of overall hypotheses (optional)

The experimenters are aware of the overall hypothesis.

15. My target sample size is (optional)

We will adopt 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. The exact design is as follows:

1. Collect data from $N_{min} = 60$ participants.
2. Compute the BF with a weakly informative prior.
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.

Participants excluded from the statistical analysis due to pre-specified exclusion criteria will be replaced by new participants and the number of exclusions will be reported.

16. The rationale for my sample size is (optional)

We quote from our manuscript (section 2.1):

"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, but 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).[2] 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.

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

- 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 analysed 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).[3] 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.

Figure 1 summarizes the results of the simulations (10,000 simulations for each combination of sample size and simulation type).[4] 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 Figure 1 (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 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)."

[3] 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).

[4] We report only simulations for which the models converged; see Appendix C for convergence failure rate.

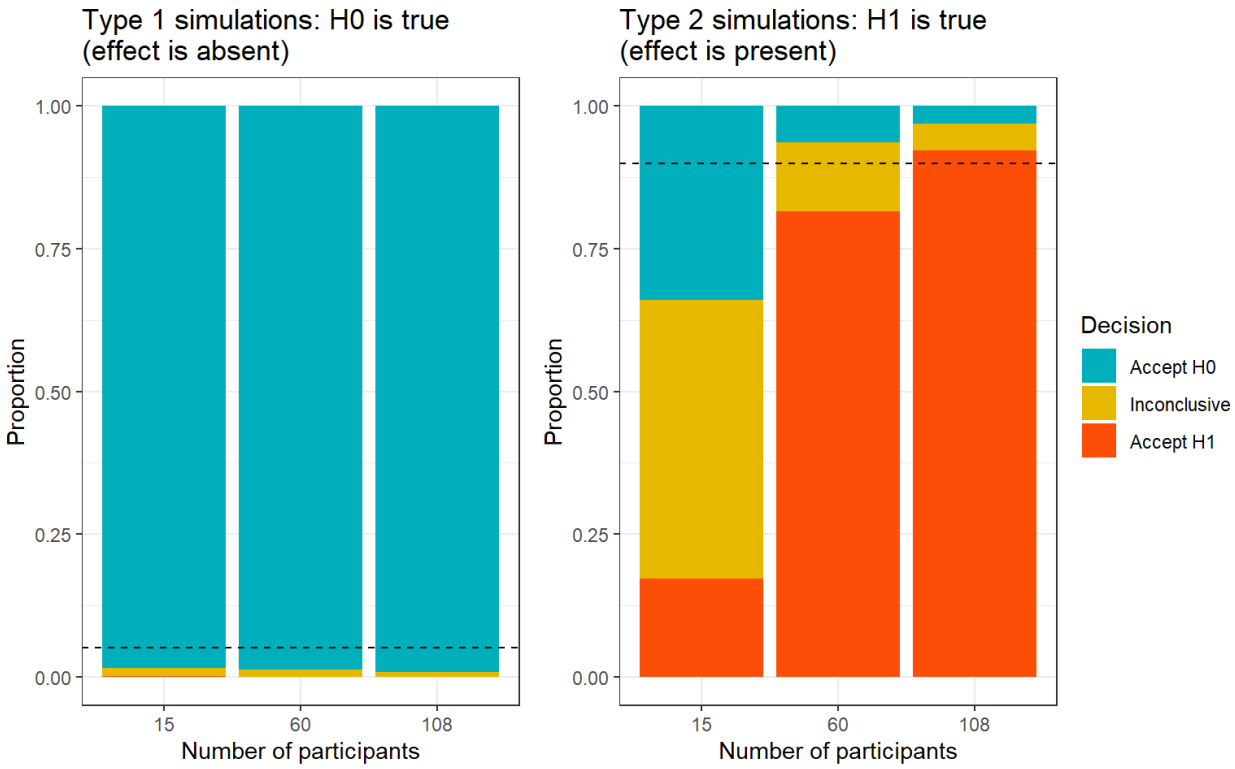

Figure 1. Summary of Bayes factor design analysis. For each simulated data set, the decision could be to either accept the H0 (if BF ≤ 1/6, blue fill), remain undecided (if 1/6 < BF < 6, yellow fill) or accept the H1 (BF ≥ 6, red fill). The plots show the proportion of decisions per sample size (x-axis) 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 10,000 simulations.

Documenting Differences between the Original and Replication Study

17. The similarities/differences in the instruction are (optional)

The instructions are as close to the original as we could make them based on the information provided. Exact instructions were not reported. We contacted the authors and asked them for the experimental protocol, but it was not available.

18. The similarities/differences in the measures are (optional)

We will collect the same raw data as SP13 (audio recordings of participants carrying out the task). However, the coding for our main dependent measure (Errors) will differ as follows. We will adopt a binary coding for the data: For each word within a 4-word memory trial, the dependent variable

will be 1 if the verb is recalled and 0 if it is not. Our coding differs from that of SP13 in that it disregards shift errors, an error type that accounted for 12% of all errors in the original and whose removal did not affect the critical interaction effect (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.

[The following passage is copied from the manuscript, sect. 3.1:]

"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 well be coded as an omission and an addition (two errors). For want of a principled protocol that can be implemented in a machine, we 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), exactly the same as if the response had been *peel-potato-garden-I don't know* (two replacement errors and an omission). Intuitively, it would seem that the former response reflects superior memory than the latter, but this would not be captured by the coding. 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.[5]"

19. The similarities/differences in the stimuli are (optional)

Our stimuli are Swedish verbs instead of English verbs. They are otherwise similar with respect to all relevant dimensions, see point 11.

[5] 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.

20. The similarities/differences in the location (e.g., lab vs. online; alone vs. in groups) are (optional)

The type of location is the same based on the information provided in the article. Our sessions will take place in a quiet room with only the participant and the experimenter.

21. The similarities/differences in remuneration are (optional)

In both cases, participants were reimbursed for their time. The exact remuneration is not specified in SP13. We will pay participants a gift voucher with a value of approximately 100 SEK (~10 EUR) per hour.

22. The similarities/differences between participant populations are (optional)

Our study will be conducted in Sweden and we will therefore recruit native speakers of Swedish (rather than native speakers of English). Because there is no theoretical motivation to suspect that the effect reported in SP13 is language-specific, this difference should not impact the results. In all other respects, the populations are as similar as possible based on the information provided by SP13, as we detail next.

Participants will be in the same age range as the original (18–30). As in the original, we will screen participants for right-handedness, normal vision and hearing, and lack of history of neurological or psychiatric illnesses. We will exclude musicians, operationalized as anybody who has at least five years of formal musical training or equivalent informal experience. We will also exclude participants who report having played the drums for more than one year.

Monolingual Swedish speakers are virtually impossible to find in the 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.

23. What differences between the original study and your study might be expected to influence the size and/or direction of the effect? (optional)

We do not expect any of the differences to influence the size and/or direction of the effect. Our increased power (due to substantially more participants and slightly more items) is only expected to increase precision of our estimates of the effect, but not to modulate the effect per se.

An additional difference not mentioned above is that we will only run the two critical conditions (movement interference with either arms or legs), and remove from the design the control and articulation conditions (SP13, p. 224-225). We choose to restrict the design to address the critical effect, namely the interaction between word type and movement. In the original study, it is this 2x2 subset of the design that is reported as the analysis "directly addressing the main hypothesis motivating this study" (SP13, p. 226). Because the order of the additional conditions was

counterbalanced following a Latin-square design (SP13, p. 225), it should not be a confound in the reported results.

In sum, there is no theoretically motivated reason to expect that the true underlying effect should differ between the original study and our replication due to this or any of the above-mentioned differences.

24. *I have taken the following steps to test whether the differences listed in the previous question will influence the outcome of my replication attempt (optional)*

We have normed the stimuli (Swedish verbs) with respect to all variables that the original authors deemed important.

Analysis and Replication Evaluation

25. *My exclusion criteria are (e.g., handling outliers, removing participants from analysis) (optional)*

At the trial level we will apply two types of exclusion criteria:

1. We will exclude trials in which the participant starts the oral recall before the beep, that is, if the word onset falls before the end of the 6 second memory phase (see **Error! Reference source not found.**).
2. In the two interference conditions, we will exclude trials in which participants fail to execute the interference task, which we define as starting the paradiddle later than 3 seconds into the memory phase (that is, if the first tap is registered later than 3 seconds after the offset of the fourth word in the trial).

At the participant level, we will exclude participants for whom either of the above criteria or technical failure (e.g., recording not working) result in excluding more than 30% of the trials across blocks or more than 50% of trials in a single block.[6] All exclusions will take place before the recall data is coded and analysed. Excluded participants will be replaced.

26. *My analysis plan is (justify differences from the original) (optional)*

Count data (in this case errors in the memory task) violate the statistical assumptions of ANOVAs and t-tests (Jaeger, 2008), which were the statistical analyses used in SP13. Instead, we will analyse the data using a Bayesian version of logistic mixed effects regression implemented in the package *brms* (Bürkner, 2017) in the R statistical environment (R Core Team, 2015). Logistic mixed effects regression is well suited to model binary outcomes and relies on the log of the odds as a link function (see Jaeger, 2008). The dependent binary variable Error (=1 if a word is missed, =0 if it is remembered; see Data coding) will be modelled as a function of the contrast-coded predictors Movement (1=arm movements, −1=leg movements), Word Type (1=arm-related words, −1=leg-

[6] 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.

related words), and their interaction. The model will include crossed random effects by participants and items. To determine the exact random effect structure of the model, we will follow the guidelines in Barr, Levy, Scheepers, and Tily (2013), fitting whenever possible the full random structure motivated by the design. Additionally we will include the following nuisance variables as fixed effect predictors in the model (centred and scaled): trial position within the experiment, error on any of the preceding words in trial (binary), word position within trial.

Statistically, our analysis plan represents an improvement over the original, as the underlying probabilistic model (logistic regression) is more appropriate to the phenomenon and the model allows for parsimoniously capturing by-participant and by-item random variability (Baayen et al., 2008; Clark, 1973).

27. A successful replication is defined as (optional)

To decide when to stop data collection (see point 15) and to make a decision as to whether our replication successfully detects the effect reported in SP13 or fails to do so, we will use Bayes factors (BF) (see Dienes, 2014; Verhagen & Wagenmakers, 2014, and references therein). We will compute the following two BFs (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).

Our decision as to when to stop data collection will be based on the calculation of BF1 only. Once data collection has stopped (either because BF1 $\geq$ 6 in favour of one of the competing hypotheses or because we have reached N_{max}=108) BF2 will be computed.

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

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

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

References

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

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Penna, N. D., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., … Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad9163

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

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

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

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

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

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

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

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

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

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

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

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review*, *23*(4), 941–958. https://doi.org/10.3758/s13423-016-1045-2

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*, 0021. https://doi.org/10.1038/s41562-016-0021

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

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

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

Sawilowsky, S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, *8*(2). https://doi.org/10.22237/jmasm/1257035100

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

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

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

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

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. (forthcoming). Making replication mainstream. *Behavioral and Brain Sciences*, 1–50. https://doi.org/10.1017/S0140525X17001972